Transparently Reported Research: An analysis of Wellcome-funding publications in 2016 and 2019

2020-Oct-28

ripetaReview

Contents

Executive Summary

This report summarises the transparency reporting practices of *Wellcome Trust*-funded research during the years of 2016 and 2019.

Using open-access, published research articles, we compared Ripeta's quality criteria across publications from each of the years. These quality criteria have been shown to improve reproducibility and are based on best practices guidelines for research transparency. The scientific research articles are analyzed across criteria important for transparently reporting research:

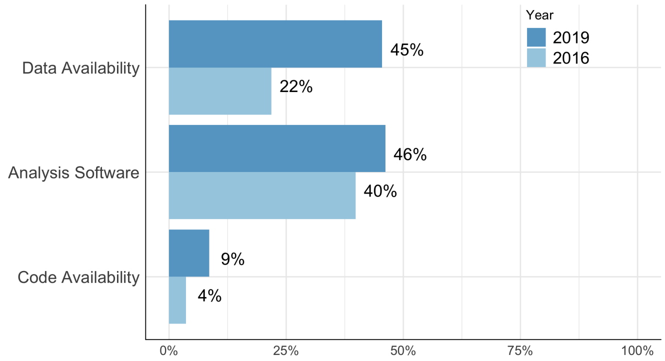
- Presence of a data availability statement.
- Location of available data.
- Data repository use.
- Reporting of analysis software.
- Reporting of analysis code.

These criteria are not uniformly implemented, making this ripetaReview an important step toward improving reproducibility and science reporting.

There are 18,187 *Wellcome*-funded articles published in 2016 (N=8,794) and 2019 (N=9,393) focused on research only (i.e., not commentaries or editorials). Of those, 93.3%(16,975) are freely available, while 6.7%(1,220) are available for a fee from the publishers.

A random sample of approximately 6,200 open-access articles - 3100 articles for each year - have been used for the analyses.This analysis shows *Wellcome*-funded authors employed some important reporting practices.

- The number of articles with data availability statements (DAS) increased 23.7% from 21.8% in 2016 to 45.5% in 2019.
- There has been little change in the number of authors sharing data in repositories or other platforms between the comparison years (36.8% in 2016 and 35.2% in 2019). The top five data sharing platforms include Github, OSF, Figshare, GEO, and GenBank.
- 39.83% in 2016 and 46.18% in 2019 listed some type of analytic software used for their statistical analysis, with the most common type being: R, SPSS, and Graphpad Prism.
- Only 15 articles in 2016 compared to 57 in 2019 met all the quality criteria such as sharing data in a repository and sharing code, making those the most likely to be fully reproducible.



Methods and Overall Summary

The Ripeta team used the Dimensions to identify the 18,187 digital object identifiers (DOIs) from *Wellcome*-funded articles published in 2016 and 2019; 16,975 of the articles were open access. Using the Fields of Research (ANZSRC FOR) and Health Research Areas categories compiled by Dimensions, we assigned each publication to a subject area. Next we checked the proportion of papers from each subject in each year examined (2016 and 2019) and took a random sample of 6200 publications. We performed checks to ensure that the sample subject areas were adequately representative of the total body of *Wellcome*-funded papers but did not ensure those over 300 papers per subject area were equally matched. There was no statistical difference between the distribution of the sampled papers compared to the total papers by research category for each year. Below is a graph of the subject areas represented in the sample.

For this report, we focused on research articles only (e.g., not commentaries or editorials) and only those articles that are freely available (i.e., CC-BY, open access). Approximately 93% of the research was published as open-access (93.6% in 2016 and 93.1% in 2019). We stratified by subjects with more than 1500 papers to sample from and randomly selected approximately 3100 of these research articles for analyses from each year for a total sample of 6198 DOIs. Due to the comparatively low number of papers in less populated subjects such as "Environmental Science" and "Societal" Ripeta slightly over sampled those subjects and slightly under sampled the subjects with greater than 1500 papers available. That sampling method resulted in a small difference in the proportions of each subject as shown below, but ensured the smaller subject areas were included.

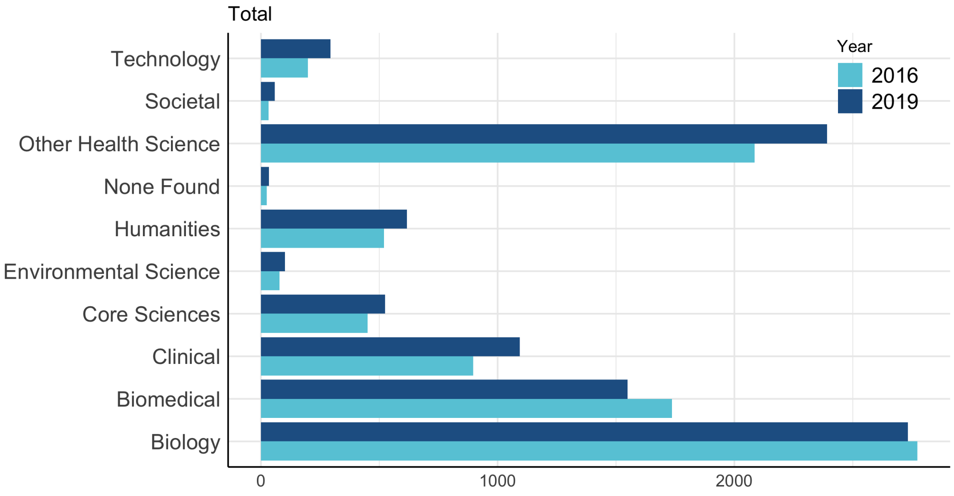


Figure 1: Total Publications by Subject Area

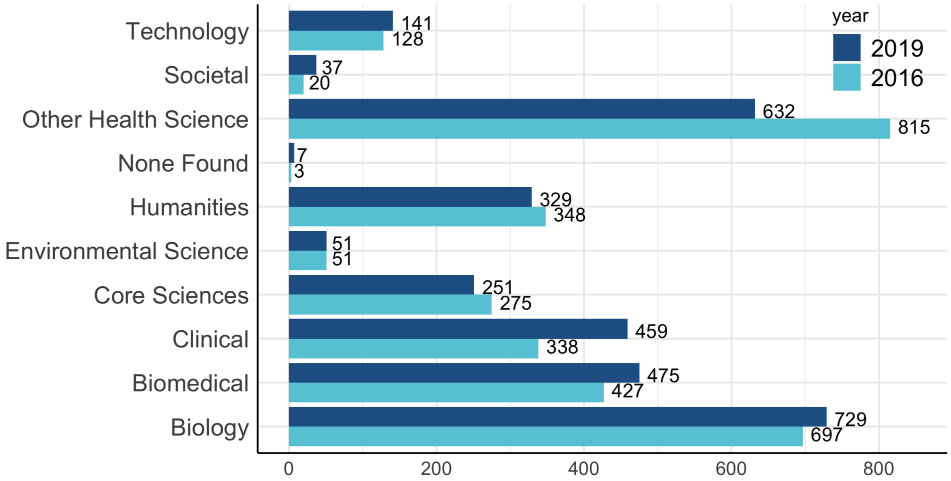
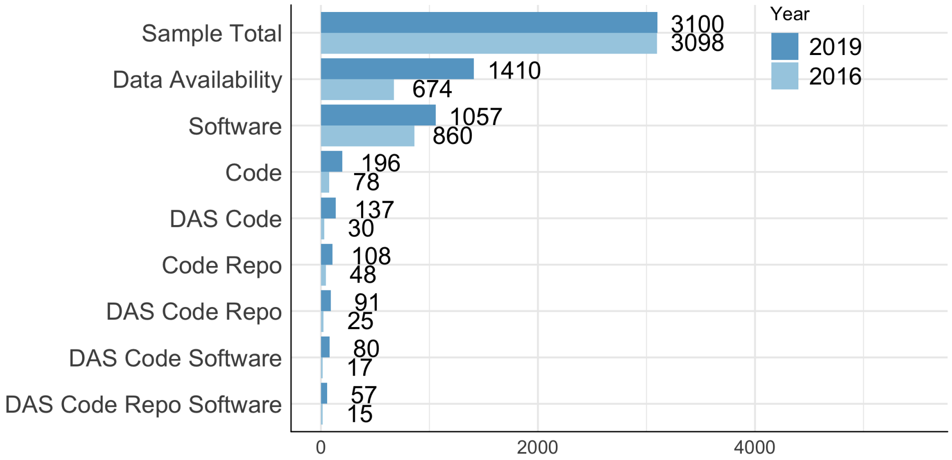


Figure 2: Sample Publications by Subject Area

We then analyzed the sample of manuscripts using the Ripeta application, which leverages natural language processing (NLP) to identify and extract key pieces of text from scientific articles. We have developed several NLP models, each tuned to a specific reproducibility criterion. Trained to read like humans, these NLP models scan articles for seed phrases and terms that indicate the presence of their respective reproducibility criteria. The NLP models can process large numbers of articles very rapidly, making them powerful tools to hold researchers accountable.

The summary table below shows (see Figure 3) the total research publications sampled from each year. The more information present, the more likely the work will be reproducible. While each category may have a small number of publications, the number of publications fulfilling all of the criteria increased noticeably between 2016 and 2019 as shown below.

Future work can help to track *Wellcome*'s progress by analysing a larger sample of articles, examining other reproducibility criteria, or employing stricter definitions of the criteria examined in this report. Additional information about our approach, technology, and definitions may be found in the appendix.


Sample Total
Data Availability
Software
Code
DAS Code
Code Repo
DAS Code Repo
DAS Code Software
DAS Code Repo Software
3100
3098
1410
674
1057
860
196
78
137
30
108
48
91
25
80
17
57
15
0
2000
4000
Year
2019
2016

Data Sharing and Data Documentation

Data availability is crucial to reproducibility. With access to raw data, researchers can: 1) Determine if the data match the conclusions of the article, or 2) Compare their own raw data to that of original study to determine the accuracy of the data. In other words, access to raw data allows researchers to identify analysis errors or falsified data. Moreover, the practice of making data available helps the original research for future work.

Wellcome guidelines emphasize the importance of data availability in its submission guidelines, which state, “We expect our researchers to maximise the availability of research data, software and materials with as few restrictions as possible.” [1] To help improve rates of reproducibility, Wellcome Trust even offers a platform for sharing data and other key elements of reproducibility. [2]

Currently we analyze two criteria related to data accessibility. First, the Ripeta application checks for the presence of data availability statements, which provide transparency to the reader and give key information about how to access data. Second, we track where and how authors make their data available. Research has shown that data location affects both ease of access and data completeness, making data location a crucial factor in data accessibility. In particular, our research has shown that data tend to be more accessible and complete when they are stored in external repositories.

This section displays the percentage of papers with data availability statements, the storage locations of the data (derived from the Data Availability Statements or Materials and Methods sections), and which repositories were most often used. Together, these metrics show if, how, and where authors who received grant funding from *Wellcome Trust* generally share data.

Data Availability Findings

In 2019, **45.5%** of articles contained a data availability statement, an increase of **+23.7%** since 2016.

Only **35.2%** of data availability statements in 2019 listed the data as being accessible in a repository, which has not changed significantly since 2016.

Many papers claimed to have the data in a supplement attached with the paper online or available upon request to the author, although contact information was not uniformly present.

Variable	2016	2019	Change
Data Availability	**21.8%**	**45.5%**	**+23.7%**

[1] https://wellcome.ac.uk/funding/guidance/data-software-materials-management-and-sharing-policy
[2] https://wellcomeopenresearch.org/

Data Location Findings

The most prevalent means of sharing data is through a 'repository'. Note, however, that this does not mean the data are easily available.

Repository is used in a broad sense and includes online repositories where researchers get data as well as other data sharing platforms (e.g. GitHub).

While the number of authors stating their data are not available has decreased 7.5%, the number of authors stating the data are available upon request has increased in the same time period.

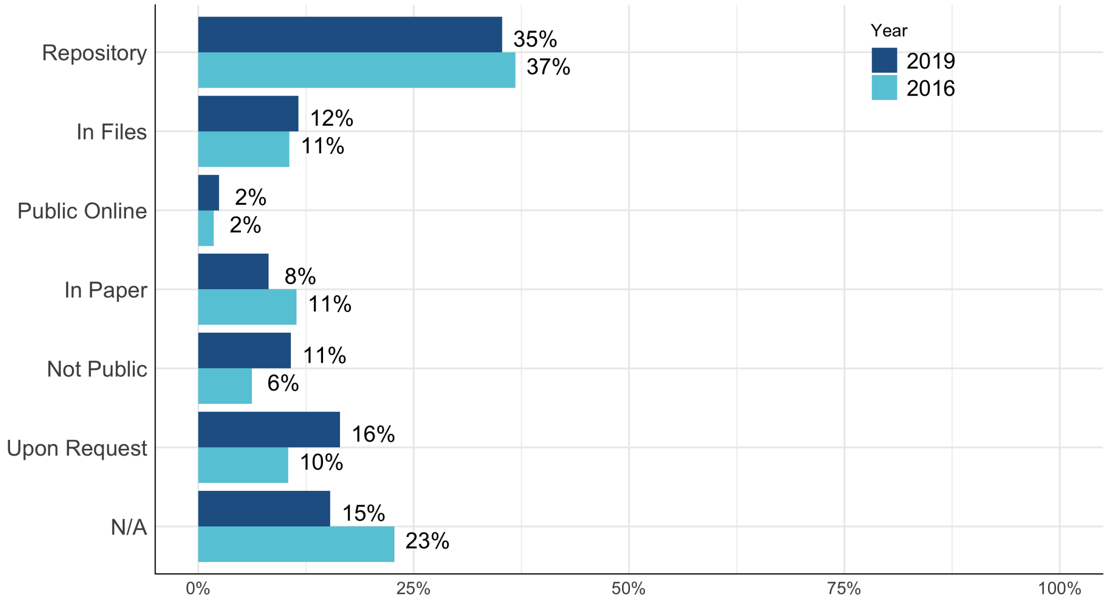


Figure 3: Summary Table of Data Locations 2016 & 2019

Data Location	2016	2019
Repository	**36.8%** (878)	**35.2%** (856)
In Files	**10.6%** (252)	**11.6%** (283)
Public Online	**1.8%** (43)	**2.4%** (59*)
In Paper	**11.4%** (272)	**8.2%** (199)
Not Public	**6.2%** (148)	**10.8%** (262)
Upon Request	**10.4%** (249)	**16.5%** (400)
Not Available	**22.8%** (543)	**15.3%** (372)

Table 1: Percentage of data sharing by location
*Derived from both DAS and Materials and Methods

Top Five Data Sharing Platforms

Researchers used over 50 repositories or data sharing platforms to host their research data, many appearing only once or twice. The majority of papers stating they shared data did not use a large repository or well-known resource, opting instead to put data in a university system or on a researcher's website.

The top five most used data sharing platforms from this sampling in both 2016 and 2019, were Github, Figshare, Open Science Framework (OSF), Gene Expression Omnibus (GEO), and GenBank. Note that the use of OSF and figshare have increased significantly since 2016 as shown in the table below.

Note: We recognize that some data cannot be widely shared because they are sensitive or contain personally identifiable information. To ensure there were not arbitrarily restricted data in these papers, we did a number of checks across the papers and found that every paper that restricted access to the data also included an ethics statement indicating that there is a valid reason for not sharing the data.

Top Platforms	**2016** (302)	**2019** (493)
Github	**16.6%** (50)	**17.2%** (85)
figshare	**7.9%** (24)	**12.6%** (72)
OSF	**3.6%** (11)	**14.0%** (69)
GEO	**10.3%** (31)	**9.3%** (46)
GenBank	**11.3%** (34)	**5.7%** (11)

Table 2: Top 5 data sharing platforms in 2016 & 2019

Data Sharing Insights & Examples

While there has been a significant increase in publications with Data Availability Statements (DAS), this has unfortunately not translated into more data being available. As shown below, a publication can be published with a DAS stating that data are either in the article or available upon request from the author. Yet, neither of those methods have shown that data are easily accessible [3]. The examples below show how authors comply with the requirement of having a DAS, however, the degree to which the data are actual available will vary. The best example has put their data in a public repository.

Example 1: Adequate
The data that support the findings of this study are available from the corresponding author upon reasonable request.

DOI: 10.1038/s41467-018-08143-4

Example 2: Better
All data generated or analysed during this study are included in this published article.

DOI: 10.1186/s12879-019-4626-7

Example 3: Best
Datasets from all human and monkey experiments, analysis code and model associated with this work are available on Dryad doi:10.5061/dryad.53sq7kn.

[3] Federer LM, Belter CW, Joubert DJ, Livinski A, Lu Y-L, Snyders LN, et al. (2018) Data sharing in PLOS ONE: An analysis of Data Availability Statements. PLoS ONE 13(5): e0194768. https://doi.org/10.1371/journal.pone.0194768

DOI: 10.7554/elife.40145

Variable	**2016** (2159)	**2019** (2293)	Trend
Software Stated	39.8% (859)	46.2% (1059)	+6.4%

Table 3: Percentage of manuscripts with software stated in 2016 & 2019

Analysis Software

Even if authors provide thorough descriptions of their analysis methods, they often render their work irreproducible by failing to state the software and code used to conduct those analyses. This is problematic as the use of different software or code could produce very different results [4]. Ideally, a software statement should cite the software used, state which version of the software was used, and in the case of open source software types state which packages/libraries were used to carry out analyses. Some statements may go even further and provide details regarding how the software was implemented such as names of specific functions used or parameters passed to those functions.

Wellcome readily acknowledges the importance of software and even calls on authors to cite their specific software versions in the following section of their author guidelines: *Wellcome expects all users of research data, software and materials to cite the source, and to abide by the terms and conditions under which they were accessed.*[5]

As shown in the following tables and graphs, there is an upward trend of 6.7% in stating the software used for analyses within publications.

Analysis Software Availability Findings

Because not all research uses statistical software, we assessed how many manuscripts we would expect to have software stated based on the type of research they describe.

Our analyses show that 73.9% (n=4448) of the total papers should have stated software.

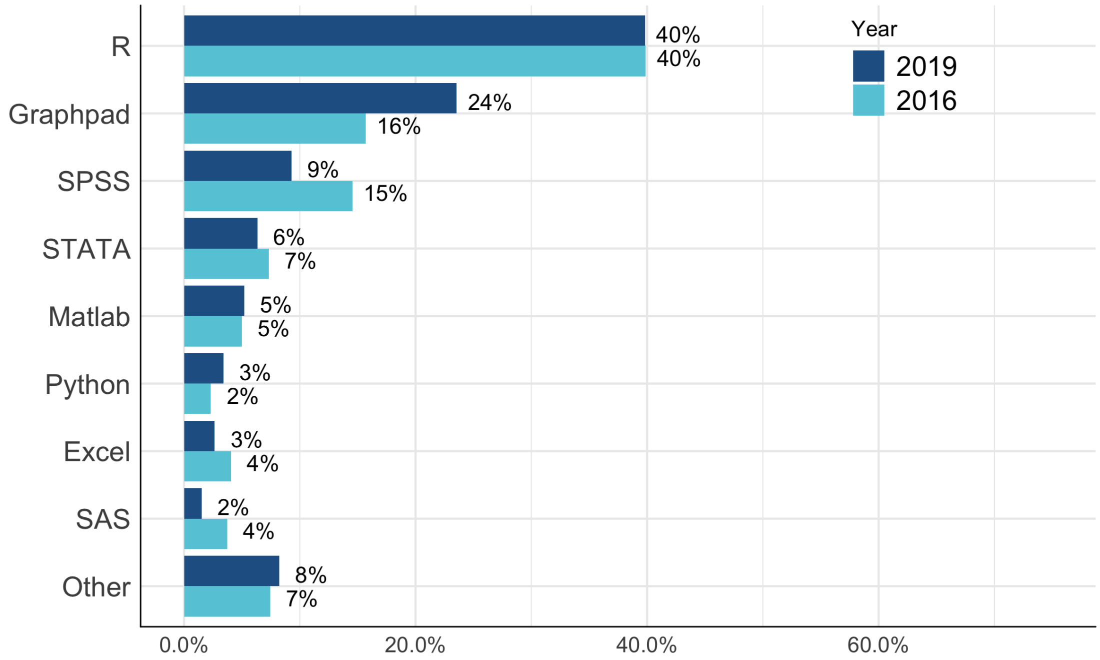
Of the papers expected to have analysis software stated, 39.8% (2016) and 46.2% (2019) stated the analytical software they used.

The most common software used for analyses were R, Graphpad (including Prism), and SPSS.

[4] Gronenschild EHBM, Habets P, Jacobs HIL, Mengelers R, Rozendaal N, et al. (2012) The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. PLOS ONE 7(6): e38234. https://doi.org/10.1371/journal.pone.0038234

[5] https://wellcome.ac.uk/funding/guidance/data-software-materials-management-and-sharing-policy

Analysis Software by Type



Software Type	2016	2019
R	**39.9%** (343)	**39.8%** (421)
Graphpad	**15.7%** (135)	**23.6%** (249)
SPSS	**14.5%** (125)	**9.3%** (98)
STATA	**7.3%** (63)	**6.3%** (67)
Matlab	**5.0%** (43)	**5.2%** (55)
Python	**2.3%** (20)	**3.4%** (36)
Excel	**4.1%** (35)	**2.7%** (28)
SAS	**3.7%** (32)	**1.5%** (16)
Other	**7.4%** (64)	**8.2%** (87)

Analysis Software Insights & Examples

Analysis Software statements were evaluated along a continuum of reproducibility to indicate the least information to the most information provided. Outside of sharing code, the more information provided, the more likely the research is to be reproducible. Thus, simply stating the software used meets a minimal requirement of transparently reporting the work, however, software should also be correctly cited. Including further details such as software version and naming specific packages, libraries, or functions used provides necessary details to potentially reproduce the computational analyses.

The following examples highlight the differences between reporting minimally necessary information and more robustly describing the research analyses.

Example	Rating	Software	Citation	Version	Pkg/Library	Function
1	Adequate	Yes	In Text	Yes	No	No
2	Better	Yes	In Text	Yes	Yes	Yes
3	Best	Yes	Yes	Yes	Linked	Linked

Example 1: Adequate
The SPSS software (version 20, IBM Corporation) was used to carry out statistical analyses. The Holm-Bonferroni procedure was used to correct for multiple comparisons. The corrected p-values (pc) of less than 0.05 were considered significant.

DOI: 10.1016/j.neurobiolaging.2018.11.009

Example 2: Better
Confidence Intervals (CIs) for Poisson counts (HBR, CBR and EIR) were estimated using the exact method of the poisson.conf.int() function in the epitools package version 05-10 of R software version 3.3. Binomial CIs were estimated for proportions (SI, EI and CBI) using the exact method of the binom.confint() function in the binom package version 1.1-1 of R software version 3.3.

DOI: 10.12688/wellcomeopenres.14761.3

Example 3: Best
Software GWAS were performed with SNPTESTv2.5.0 [47] and meta-analysed with META v1.7.0 [48]. The ‘qqman’ R package [http://cran.r-project.org/web/packages/qqman/] was used to create Manhattan/quantile-quantile (QQ) plots. The ‘bedtools’ ‘clusterBed’ function [https://github.com/arq5x/bedtools2][49]was used to group SNPs associated at p 1e-06 (except for the ’Meta-Analysis 2’, see footnote of Table 2) into 1Mb clusters, which were then annotated with ANNOVAR [50] [http://annovar.openbioinformatics.org/en/latest/]. Regional association plots were produced with LocusZoom v1.3, using an hg19 reference (1000G March 2012)[51].

DOI: 10.1186/s40246-018-0190-2

Analysis Code Availability

The Analysis Code availability variable gauges whether or not authors have shared their analysis code. Sharing code used during a study ensures the work is as transparent and reproducible as possible. Even if software is properly cited there may be different ways to implement that software which change the results of analyses either slightly or significantly. Code sharing is a crucial counterpart to data sharing and can very quickly allow other researchers to verify and expand upon a work, or test the analysis themselves. Because of this, sharing code is an focal component of transparent, responsible science. A lack of understanding of the data, analysis software and specific code used greatly hinders the ability to reproduce an article's processes and findings.

Analysis Code Availability Findings

A subset of the original set of papers would be expected to used code. Based on the types of research, we expected 2159 papers in 2016 and 2293 papers in 2019 to include code.

There is a 5.0% upward trend in sharing code within publications.

Variable	2016	2019	Trend
Code Shared	3.6% (78)	8.6% (196)	+5.0%

Analysis Code Insights & Examples

Analysis Code statements were also evaluated along a continuum of reproducibility based on how accessible the source of the code listed is. The least reproducible code statements may ask interested parties to contact the author who will then share analysis code with them. 'Better' statements may have code deposited in Github, or another common online source. 'Best' code sharing statements will provide detail the code available and will indicate that authors have deposited the code in an interactive environment conducive to easy reproducibility such as Gigantum, Jupyter Notebook sharing platforms, or Code Ocean.

Example 1: Adequate

All supporting data, code and protocols have been provided within the article or through supplementary data files.

DOI:10.1099/mgen.0.000083

Example 2: Better

The accession numbers for the raw sequencing data reported in this paper are GEO: GSE137710 and GEO: GSE130201. Scripts reproducing the analysis will be available on request.

DOI:10.1016/j.cell.2019.09.035

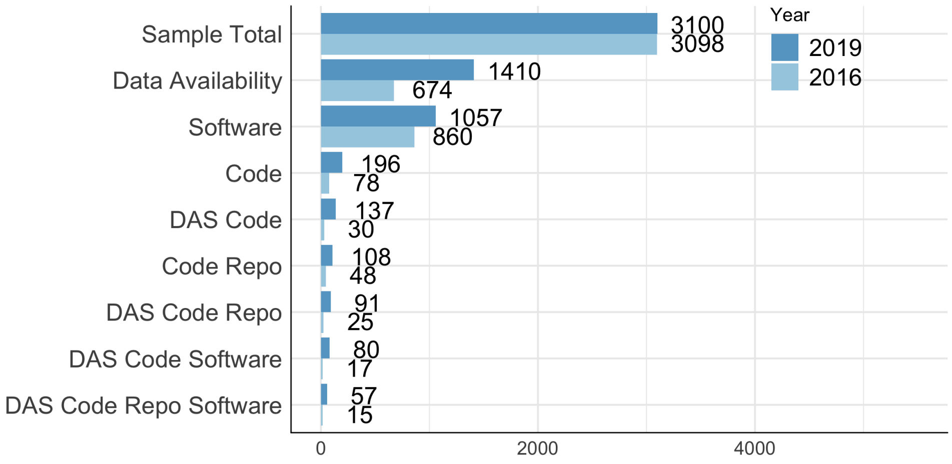
Example 3: Best

The data analysis procedures for the CGM were implemented in R using the additional packages RMySQL, outliers, matlab, amap, RSvgDevice and RSvgTipsDevice. We developed ChemGRID (http://chemgrid.org/cgm) as a webportal for the upload, processing and visualisation of chemical-genetic screen data (Fig. 4a,b) using PHP, PEAR, Perl and MySQL. The cheminformatics functionalities to register structural information were implemented with Python and FROWNS, PerlMol and MolDB4 (ref. 22). All code for data analysis is available at https://github.com/jwildenhain/chemical-genetic-matrix.

DOI:10.1038/sdata.2016.95

Conclusion & Recommendations

Overall, Wellcome is doing quite well in supporting and advancing open research principles. As the chart below shows, they have seen an increase in data availability statements and software and code sharing.



Overall Recommendations:

Clearly define what should be shared and how the research outputs should be prepared for potential reuse and verification

For example, develop guidance and a set of expectations for how code, scripts, data, protocols, and more should be managed, curated, and shared when appropriate

Provide training and support on these expectations - or partner with local institutions or organizations to provide training and support (e.g. ripetaTraining)

Consider funding the quality checks similar to funding APC fees.

Pay particular attention to LMIC researchers who may need supplemental support for any automated checks.

Offer authors a means to check their manuscript before it is submitted to a journal or on Wellcome Open Research to offer suggestions for improvement. (e.g., ripetaReview)

Implement a research quality check into the grant or contract proposal review stage (e.g. ripetaReview)

Specific Quality Check Recommendations:

Provide clear guidance and set expectations for how data should be shared and where. For example, develop or reuse an existing repository selection tool for different disciplines or data types or institutional/organizational needs.

Develop or reuse existing guidance on how to best de-identify or anonymize data or research outputs

Supporting code availability will advance reproducibility but is not required by most publishers. As a funder, you have the opportunity to support this initiative.

While not verified in our checks, there are statistical software that is more commonly used by different professionals. This report suggests that many researchers did not use a statistician on their team. Depending on the research, you may want to ask researchers if this is the case.

Contact

For more information about this report or about Ripeta, please contact us at info@ripeta.com.
All data for this report have been made available to Wellcome.

Acknowledgements

Leslie D. McIntosh, PhD
ORCID: 0000-0002-3507-7468

Dr. McIntosh is the founder and CEO of Ripeta, a company formed to improve scientific research quality and reproducibility. Now part of Digital Science, the company leads efforts in rapidly assessing scientific research to make better science easier. She served as the inaugural executive director for the US region of the Research Data Alliance and is still very active with the RDA.
She holds a Masters and PhD in Public Health with concentrations in Biostatistics and Epidemiology from Saint Louis University and a Certificate in Women's Leadership Forum from Washington University Olin's School of Business.

Cynthia Hudson Vitale, MLIS
ORCID: 0000-0001-5581-5678

Ms. Hudson Vitale is the COO and a co-founder of Ripeta. She is also the Head of Research Informatics and Publishing at the Pennsylvania State University Libraries where she leads a team focused on enhancing the technology-driven research capacity of faculty and students particularly in the areas of statistics, digital humanities, research data services, open-access publishing, and maps & GIS. She is a founding member of the Data Curation Network, a multi-institutional initiative building a shared staffing model for making research data FAIR and the principal investigator on a project developing specialized data curation training. She received her BA from St. Louis University and her MA from the University of Missouri, Columbia. Her areas of expertise include information architecture, data curation, and scholarly communication.

Josh Q. Sumner
ORCID: 0000-0002-3399-9063

Mr. Sumner is a data scientist at Ripeta using R to assist in analyzing data to evaluate research. He graduated from Appalachian State University in May of 2019 with a BS in Strength and Conditioning and a minor in Statistics. Josh is now pursuing an MS in biostatistics at Washington University in St. Louis.

Ripeta Team
This report is made possible through a team effort of Ripeta data scientists and engineers including: Chris Westling, Ruth Whittam, Paolo Ortega Saborío, and Leah Haynes.

Appendix

The Ripeta application uses natural language processing (NLP) to pull key pieces of text from an article. The NLP was trained to read like a human, scanning articles for seed phrases and terms that indicate the presence of one of the reproducibility criteria. These seed phrases and terms were identified by the Ripeta team, who manually annotated over 500 papers. During these annotations, team members made note of language patterns that signaled the presence of one of the criteria, as well as the typical location of those criteria within the paper. Next, the Ripeta team developed a spaCy model that used these seed terms to start recognizing language patterns. After the Ripeta application scanned its initial batch of articles, the Ripeta team refined its seed terms to improve accuracy. This cycle continued until the Ripeta application could consistently produce accurate results without human interference or review.

Ripeta Criteria Definitions

Data Availability Statement

Definition: A statement, in an individual section offset from the main body of text, that explains how or if one can access a study's data. The title of the section may vary, but it must explicitly mention data; it is therefore distinct from a supplementary materials section.

Data sharing is fundamental to research reproducibility because data allows other researchers to determine if the stated analysis methods and data could have actually produced the stated conclusions. In other words, data sharing allows other researchers to identify the presence of falsified data or faulty analysis methods. The research community widely accepts the benefits of data sharing and has attempted to hold authors accountable by requiring data availability statements. These statements, when properly offset from the main text, provide readers with a quick way to determine how to access data.

Data Location

Definition: Where the author states that the article's data can be accessed, either raw or processed.

Even when articles have data availability statements, their data are not always easily accessible. Where and how authors make their data available strongly influences how easy they are to access. Furthermore, research shows that data location can also serve as a proxy for the completeness of the data; for instance, full data sets are more likely to be available when they are shared in external repositories or upon request rather than when they are made available in the article or supplemental files. Therefore, Ripeta believes it is very important to track how authors make their data available, not just whether or not authors have included data availability statements. Particularly, we see if authors have stored their data in external repositories, as data tend to be both easier to access and more complete when they are stored in this way. Currently, we have ranked locations in the following order:1) external repository, 2) in article and supporting files, 3) online, 4) within paper, 5) not publicly available, 6) upon request, and 7) not applicable.

Ripeta recognizes that there are legitimate reasons for authors to restrict access to certain data, particularly protected health information. We are therefore training our software to identify when authors have stated reasons for their data restrictions. This improvement will allow us to differentiate between legitimate and arbitrary data restrictions, and we will be able to rank papers more highly when they have provided a reason for their restrictions.

Importantly, Ripeta only checks the author's report of where data can be found; we do not currently check whether data can actually be accessed or are sufficient to reproduce the findings.

Analysis Methods Stated

Definition: The article includes an explanation of the methods used for analysis, including statistical analysis.

Authors routinely provide thorough descriptions of the methods used to conduct their research, but many authors neglect to fully describe their analysis methods. Yet data analysis is how researchers pull their data together to create conclusions, essentially the moment of knowledge creation. Therefore, analysis methods must be both robust and appropriate for the results to hold. So, when authors fail to adequately describe their analysis methods, they leave room to question the quality of their work. Other researchers will not be able to determine if the analysis methods were appropriate or possibly mischaracterized. Furthermore, without analysis methods, it would be difficult to tell if data were falsified or misconstrued. Finally, because the analysis methods affect the results, no other researcher would be able to accurately reproduce the work.

Note, Ripeta currently only checks for an explanation of analysis methods, not the quality of that analysis. Many descriptions of analysis methods are not robust enough for full research replication.

Analysis Software Stated

Definition: Authors have indicated the specific software they used to conduct their analyses.

The analysis methods determine what conclusions are drawn from the data, and software can deeply influence the analysis process. When authors fail to state what software they used, they render their work unreproducible. We analyze papers for the presence of software information, as well as the specific software that authors report using.

Code Shared

Definition: Authors have shared access to the most updated code that they used in their study, including code used for analysis.

For many studies, code is a critical piece of the data analysis process. Yet most authors do not share their code, making it difficult to properly reproduce their studies. Many publishers recognize the importance of code sharing and require it in their author guidelines, but Ripeta has found that compliance is typically below 10 percent. Thus, Ripeta checks if authors have shared access to their code, preferably in an external repository.

Data Location Definitions

Data location buckets are listed below by order of preference, as in how we would prefer data be shared. For the latter three ('not publicly available,' 'upon request,' and 'not applicable') we are considering sub-buckets of 'reason given' as opposed to 'no reason' since justification of why data may not be publicly available for example could be significant. There are examples for these sub-buckets below but note that we are not officially using them yet.

External repository

- Data location includes keywords of "deposited" and "database".
- Gives accession number and/or link.

Example:

- All raw sequencing data and ancillary analyses are deposited in the GEO database under the accession number GSE94518.
- Crystallographic data have been deposited in the Cambridge Structural Database as specified above for each compound under the deposition numbers CCDC 1547761-1547775 that is available from The Cambridge Crystallographic Data Centre via ¡¡url¿¿.
- All data files are available from the Dryad database (doi:10.5061/dryad.q3b16).

In article and supporting files

- Contains a statement along the lines of "all relevant data can be found. . . ."
- "...in published article (and its Supplementary. Information files)"
- "...within the paper and its Supporting Information files."
- This is different from the "within paper" bucket (explanation below).

Example:

- All data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

Online

- This bucket is harder to identify. Will say something along the lines of "further information/data/files/etc. are available on [journal's] website" or "data sets are available at [url]."
- For the latter case of stating data is online, we will have to differentiate between urls linking to a database (in which case the bucket would be "external repository") and an example such as the second listed above.

Examples:

- All statistical data are deposited in Supplementary Table 1, which is available at the journal's website.
- Public use data sets are available at https://www.icpsr.umich.edu/icpsrweb /ICPSR/series/253.

Within paper

- Similar to "In article & supplementary files" but differs in that it does not include supplementary files and is thus a less robust data sharing method.
- Most cases of "Within paper" will be very similar to the examples above.

Examples:

- All relevant data are within the paper.
- All data generated or analyzed during this study are included in this published article.

Not publicly available

- Not publicly available data sharing is very consistently stated as in the two examples above following the pattern of "datasets (generated and/or) analyzed during/in the current study are not publicly available due to. . . ."
- If lacking the "due to. . . " part of the statement, the data location would fall into the "without reason" sub-bucket.
- We divide the last 3 buckets ('not publicly available,' 'upon request,' and 'not applicable') into "reason given" and "no reason" as it is preferred if data is not shared that there is an ethical/valid explanation given.

Examples:

- The datasets generated and/or analyzed during the current study are not publicly available to preserve the confidentiality of the respondents.
- The datasets analysed are not publicly available due to the ongoing nature of these implementation efforts and difficulty removing identifiers.
- No reason
- Data is not publicly available.

Upon request

- "Upon request" will always contain the phrase "available upon request" or "upon/on reasonable request"

- Data locations in the bucket “Upon request (reason given)” may coincide with the bucket of “Not publicly available”

Examples:

- Due to ethical restrictions, the raw data underlying this paper are available upon request to the corresponding author.
- The Atherothrombotic Dataset is not publicly available due to additional analyses that are ongoing and current but are available from Dr. DeFilippis on reasonable request.
- No reason:
  - All relevant data are available from the authors upon reasonable request.

Not applicable (N/A)

- N/A is certainly the least favorable of buckets.
- It is rare that a reason is given, but we have the sub-bucket of “reason given” since in the off-chance that there is a reason, it is significant.
- A data location will be in this bucket simply if it says not applicable (N/A).
- Often a DAS may have a heading and put N/A under since the journal requires a DAS.

Examples:

- Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study
- No reason:
  - Availability of data and materials Not applicable