Position Statement: Data and Physiology

1. Executive summary

Issues around the acquisition, analysis and publication of data are part and parcel of science, but while these debates traditionally remained in the academic realm, they are now spilling out and affecting public trust in research and the reporting of modern science, not helped by hyperbolic press releases and media articles. At the same time, analytical capabilities are expanding and the ability to study massive datasets is unlocking new frontiers of understanding and predictive capabilities. Publications in physiology often rest on real-world data collected from animals and humans, with all the variability and uncertainty this entails.

In order for experimental conclusions to be solid and reproducible, publication should include a full report of data gathering including *a priori* experimental design, uncertainty and caveats. Analysis should be detailed in full with justification for the methods utilised. Full experimental data should be made available to readers as a matter of course, unless exceptional reasons can be justified. New technologies and online hosting services mean this should also apply in Big Data studies.

Journals have a responsibility to ensure they have the capability to review the accurate and effective use and presentation of data in submitted manuscripts, and should use editorial policies to promote unrestricted access to experimental data. The skills of editors and reviewers should be supplemented by statistical experts to balance the reviewing workload.

The generation of knowledge through scientific efforts is fruitless unless this knowledge is accessed and acted upon. Policymakers have long been encouraged to base their decisions on available evidence, but these decisions should also be able to be scrutinised. When providing evidence or expertise to policymakers, they should be encouraged to give a transparent summary of the evidence used, in line with the recommendations of Sense About Science.

2. Recommendations

- Early-career researchers should receive training in working and publishing with reproducibility as a primary concern in their experimental design and write-up.
- Accuracy of data collection should be considered in *a priori* experimental design, and appropriate sample size, determined by appropriate statistical methods, must be used to ensure conclusions are valid. These considerations should be discussed in the published study.
- All analytical and statistical techniques used on experimental data should be justified by the researcher as the most appropriate technique and decided upon before the experiments are performed.
- Any data used to generate a predictive scale or algorithm should be robustly demonstrated to give sufficiently accurate predictive power using an appropriate number of variables.
- Any project data in a specialist field for which a dedicated data repository exists should be uploaded as a condition of publication.

- Uploading of massive datasets to a service such as Figshare, or a publicly available repository, should be integrated into the publishing process, making it a default step unless the authors and journal Editor agree an exemption.
- Journals should ensure the reviewing process includes the expert capacity to check statistical accuracy in all appropriate cases.
- Funding organisations should, as standard, require experimental data from studies they support to be made available to readers of the relevant article(s), or for the publication to include a statement clearly giving reasons for any inaccessible data.
- When using scientific data as part of a policy evidence base, policymakers must be strongly encouraged to follow the Sense About Science framework for transparency in evidence use.

3. Introduction

Research in any discipline has always been data-driven, but with the expanding technological capabilities for generating, storing and analysing data it has never been such a powerful tool. This in turn has led to a greater accessibility of complex tools and a greater depth of analysis being the new normal in published studies. This is the case in many branches of physiology, with more accurate data gathering and more complex data analysis available than ever before. It is important that researchers fully understand the methods they use on their data, and can explain and justify the conclusions they draw from it. There is a high level of scrutiny applied to research data, both in the course of publication and peer review, and subsequently if independent replication of the study is attempted. This statement aims to provide The Society's Members, and other scientists, with an understanding of best practice concerning generation, publication and accessibility of research data. This is the standard they should aim to achieve, and promote to others. This may include those outside of research who go on to use the data and conclusions in policymaking or other contexts.

Though the current buzzword of "Big Data" is not well-defined or comprehensively understood, the results of applying powerful analytical techniques are becoming transformative in many, if not all, research disciplines. Within physiology such analyses have made many contributions, allowing the development of "omics" fields and the analysis of ever-larger samples and cohorts.

Increased attention is being paid to the use of data in science, with reproducibility becoming the buzzword in an attempt to drive up quality along with the rise in quantity of the research being published. To this end, data and statistical analysis must be robust and well-chosen in order to achieve best practice and impactful research.

A number of scandals and retractions have raised concerns that published work is not as robust and accurate as once assumed, both in the media and in popular attitudes to science. One way to improve the trust in published work is to allow access to the study's raw data, allowing anyone to check the conclusions and evaluate any assumptions made by the authors. However, this is still not an automatic or natural step. The growing popularity of preprint servers, which can act as pre-submission/publication peer review, has not seemed encourage the widespread uptake of data sharing.

4. Definitions

The following terms are widely used in the discussion of data practices within science, and apply to many of the topics discussed within this position statement.

Algorithm: a rule or set of rules for performing an action or calculation on data and generating a result from it.

Big Data: usually used to mean datasets so big and/or complex that traditional processing applications are inadequate to handle them, and they could not be incorporated into a traditional journal paper format.

Machine learning: allowing computers to automatically optimise their data analysis without reprogramming each time, to generate a "best" solution according to given principles.

Open data: data that is publicly available without restriction – similar in this context to "open science", which calls for all research data and publications to be available to anyone.

Physiological computing: interfacing with technology through constant measurement of physiological parameters (e.g. eye-tracking, heartrate, etc).

Reproducibility: the published experimental details, data and analysis of a study being sufficient for an independent research group to repeat the study and achieve the same conclusions.

5. Data Acquisition

Data acquisition methods used in any experiment should always be clearly explained, along with analysis/discussion of the accuracy of data acquisition and any errors that may be encountered. This is an issue in laboratory-based experiments, but even more so with real-world data collection, especially if using non-intrusive and/or user-operated devices e.g. fitness/activity tracking, or in citizen science projects. If working with a cohort on long-term studies, inconsistencies in data collection must be understood and accounted for – this includes improvements in data collection with developing technology etc.

Accuracy of data collection must be a consideration when designing studies, and especially when considering the number of participants required, be they human or animal. With respect to research involving animals, this is in accordance with the principle of Reduction, one of the Three Rs governing animal research,[1] which mandates ensuring an experiment is being performed with sufficient sample numbers and data accuracy to generate a statistically valid conclusion. As the National Centre for the 3Rs states, "*If a study is not designed to yield robust results and publications are not reported with enough detail, the animals and research resources used in that study are wasted. Appropriate experimental design and statistical analysis techniques are key means of minimising the use of animals in research.*"[2] Similar principles can be applied to all studies, using animals or humans.

---

[1] http://www.understandinganimalresearch.org.uk/animals/three-rs/
[2] https://www.nc3rs.org.uk/experimental-design

Recommendation

- Accuracy of data collection must be considered in *a priori* experimental design, and appropriate sample sizes, determined by appropriate statistical methods, must be used to ensure conclusions are valid. These considerations should be discussed in the published study.

6. Analysis and Prediction

Analysis must be transparent, and based on protocols set out within the initial discussion of a new experimental hypothesis to prevent p-hacking (false positive reporting after analysing data for any statistically significant correlation and fitting a hypothesis around it). Researchers must be able to defend their methods, including choice of statistical techniques, any excluded results, etc. In the case where measurements are being made to give a proxy for another physiological variable, the relation, adequacy and specificity of the proxy must be demonstrated robustly.

Many human physiological studies have followed cohorts reconvened over many years, resulting in a large body of data collected. In order to achieve maximum value from these activities, researchers should consider allowing appropriate open access to these data for different groups to perform different analyses.

Frequently, experimental data is used to derive a predictive test or scale for future cases. Care must be taken over understanding the range of "normal" values among a diverse sample set such as humans' physiological data. It may be that one single variable is not enough to derive accurate predictive ability, and must be cross-referenced with a second parameter related to the issue under investigation. Similarly, data taken at only one time-point may be ineffective in giving a predictive capability, with a longitudinal investigation required to show a rate of change. This is of considerable importance in phenotyping studies conducted with the aim of clinical application for the data gathered.

The importance of this point is multiplied if the analysis and/or prediction are to be done algorithmically (covering the field of "machine learning"), with limited human oversight. The "garbage in, garbage out" principle comes into play and conclusions must be validated before any application.

Recommendations

- All analytical and statistical techniques used on experimental data must be justified by the researcher as the most appropriate technique and decided upon before the experiments are performed.
- Any data used to generate a predictive scale or algorithm must be robustly demonstrated to give sufficiently accurate predictive power using an appropriate number of variables.

7. Reproducibility

Physiology is far from immune from issues of reproducibility of published research. Many techniques that form the standard toolkit of physiology research have inherent difficulties in acquiring

consistent data, which in turn result in difficulties in reproducing the results of published studies.[3] This issue is compounded if the published paper includes incomplete reporting of experimental protocols, or selective reporting of the data obtained. The journal *Nature* describes a "*'minimal data set' necessary to interpret, replicate and build on the findings reported in [a] paper.*"[4] This should be made available as a matter of course, to allow other researchers to effectively interpret the published work. If this basic minimum cannot be achieved then researchers should be prepared to explain and justify why their data are not accessible.

The American Physiological Society (APS) has released a series of articles called Cores of Reproducibility in Physiology (CORP).[5] In a paper describing the initiative, author (and President of the APS 2010-11) Peter Wagner states "*It is not a problem per se for two purportedly identical studies to disagree, because if both are adequately described, the reason(s) for discrepancies can usually be found, and this alone may considerably advance the field. It is when the outcomes are in doubt because of poor descriptions or statistical errors that we waste our time, resources, and dollars and put future research (and even clinical care) at risk.*"[6] This neatly encapsulates the necessity for every researcher considering their work not just in terms of the scientific conclusions, but in every aspect of the reliability of the work and the justification for its informing future research. This mindset should be inculcated into every new researcher and be required by every journal editor.

Recommendation

- Early-career researchers should receive training in working and publishing with reproducibility as a primary concern in their experimental design and write-up.

8. Big Data

The field of Big Data is arguably in its infancy, but provides powerful tools and is already generating new insights. Analysis using Big Data techniques must meet the same standards as traditional statistical analysis, meaning the techniques must be valid and the integrity of the data clearly described. When it comes to physiological data, there may be additional issues compared to other sensor-based Big Data sources, such as variation in signal quality, differences in the benchmark of a "normal" response,[7] and unperceived variability in related parameters affecting the data being collected.

More investigation should be done into the applicability of Big Data analysis to physiology and where it can provide the most useful insight. Some population studies have been undertaken to make available biometric and imaging data of huge cohorts, such as the UK Biobank.[8] It is imperative that research funders recognise the need for these centrally-gathered and stored data programmes, as their capacity for information acquisition and analysis is far greater than any one research project could achieve. High-throughput phenotyping has also been conducted on mouse strains used in

---

[3] http://thenode.biologists.com/reproducibility-pathological-perspective/research/
[4] https://www.nature.com/news/announcement-where-are-the-data-1.20541
[5] https://www.physiology.org/doi/pdf/10.1152/japplphysiol.01046.2016
[6] ibid
[7] https://www.fda.gov/downloads/NewsEvents/MeetingsConferencesWorkshops/UCM575543.pdf
[8] http://www.ukbiobank.ac.uk/about-biobank-uk/

research,[9] and efforts of this type can bring a valuable data resource to the research community. It is worth ensuring that these projects, and other data sources such as NHS patient datasets, are exploited as freely as possible from a research perspective. It is also important that the volume of data is not a disincentive to research.

Publication of Big Data studies may not be able to include the experimental data due to size – best practice as suggested by The Physiological Society's journals is to upload the data to a not-for-profit data repository in order for others to be able to evaluate the reproducibility of the study. A free pilot study of this was arranged with The Society's journal publisher Wiley, using Figshare to upload massive datasets. Take-up was exceedingly low, suggesting greater incentives or stronger rules are required to promote this behaviour.

Recommendation

- Uploading of datasets to a service such as Figshare, or a publicly available repository, should be integrated into the publishing process, making it a default step unless the authors and journal editor agree an exemption.

9. Publishing Data

Various scientific fields have been seeing a spate of paper retractions in recent years, leading some to claim there is a "*reproducibility crisis*".[10] At the heart of this issue lies poor use of statistical tools and misuse of data, whether due to bad intention, lack of care or limited ability. Journals must be able to verify the appropriateness and application of the statistics used in articles, and must make clear the standards they expect in order for a study to be publishable. This is the case currently with The Society's journals, which publish guidelines on best practice when using and reporting statistics,[11] but these are not currently compulsory. The Society's journals are working towards mandating that good statistical analysis and reporting principles are adhered to, and will ask authors to include a summary of available datasets and the statistical methods used.

Problems concerning published statistics are not just relevant to authors, but to editors and reviewers as well. Increasing publication rates and greater scrutiny mean the workload facing volunteer academics is expanding, perhaps to a point where the ability or willingness of reviewers to contribute sufficient time is exceeded. Journals must ensure they maintain adequate and homogeneous standards concerning the use of data in their publications, and be confident in the rigour of reviews, especially in instances of the use of new techniques including Big Data. The Society's journals also work with a Statistics Editor, who is involved in the review process of submitted manuscripts to ensure statistical accuracy and appropriateness. This practice is recommended for other journals, as it allows reviewers to focus on the hypothesis and the work reported with confidence in the figures and data analysis. In an ideal world this stage would be part of the review process for all manuscripts, but this would be time and resource-intensive. Editors and reviewers should nonetheless be encouraged to flag submissions for statistical checks where appropriate.

---

[9] Maier *et al*, Current Opinion in Systems Biology, Volume 4, August 2017, 97-104
[10] http://www.slate.com/articles/health_and_science/future_tense/2016/04/biomedicine_facing_a_worse_replication_crisis_than_the_one_plaguing_psychology.html
[11] https://jp.msubmit.net/cgi-bin/main.plex?form_type=display_requirements#stats

It is becoming more common for research funders to mandate specific protocols around research data when publishing studies they have funded. UK Research and Innovation (UKRI) requires articles to include a statement on where the source data can be accessed, or giving reasons why it is not available. Some journals, such as those in the *Nature* group, also require authors to make data available, and detail any restrictions on this availability to editors and within their manuscript.[12] Specialised data repositories exist for specific types of data across many academic fields, and a number of these are relevant to physiology. Examples include the European Genome-phenome Archive (EGA) for the reporting of linked genotype and phenotype data,[13] and a number of journals make data submission to these public repositories mandatory. The Society supports these publishing requirements, as they both increase the public availability of data for future researchers to build upon, and encourage good practice and transparency as the benchmark requirements to achieve publication. The Society journals also endorse the Animal Research: Publication of *In Vivo* Research (ARRIVE) guidelines[14] as a checklist for effective reporting of *in vivo* experiments. Many of the principles in the ARRIVE guidelines are entirely transferrable to other aspects of research, and the advice therein should be considered an outline of effective and responsible data publishing.

Recommendations

- Journals should ensure the reviewing process includes the expert capacity to check statistical accuracy in all appropriate cases.
- Funding organisations should, as standard, require experimental data from studies they support to be made available to readers of the relevant article(s), or for the publication to include a statement clearly giving reasons for any inaccessible data.
- Any project data in a specialist field for which a dedicated data repository exists should be uploaded as a condition of publication.

10. Using Published Data

Addressing the above issues should result in reliable analysis and accurate data being put into the public domain via scientific literature. However, the responsible use of data is a requirement of future users as well as the author. In the policy sphere, campaigns are currently active in encouraging government to disclose sources of evidence and the methods used in reaching policy decisions, notably Sense About Science's Transparency of Evidence framework.[15] Transparent discussion of evidence sources and their contribution to eventual policy decisions is to be strongly encouraged. External consultation as part of the policymaking process is expected nowadays, but this should not be a one-way process, and external bodies should expect to see an acknowledgement of the evidence they submit and its consideration within the policy development cycle.

While it is accepted that decisions will sometimes be made on factors other than evidence, it is incumbent on the policymaker to acknowledge the evidence base, any weaknesses or disagreements

---

[12] https://www.nature.com/authors/policies/availability.html
[13] https://www.ebi.ac.uk/ega/
[14] https://www.nc3rs.org.uk/sites/default/files/documents/Guidelines/NC3Rs%20ARRIVE%20Guidelines%202013.pdf
[15] http://senseaboutscience.org/activities/transparency-evidence/

within it, and either how it has fed into a decision or what competing factors have been felt to outweigh its conclusions.

Recommendation

- When using scientific data as part of a policy evidence base, policymakers must be strongly encouraged to follow the Sense About Science framework for transparency in evidence use.