Response to reviewer comments

Dear Dr. Callaghan,

Thank you for considering our commentary article "Incentives are the biggest barriers to widespread data sharing" for publication in Patterns.

We thank the reviewers for time and detailed comments. We have addressed comments from the reviewers and believe the paper is clearer and better as a result. Our response to individual reviewer comments are presented in the table on page 2 of this letter.

Regards,

Nicholas Tierney & Karthik Ram

ID	comment	response
R1-1	The authors have worked to clarify the type of data that the article is geared toward. However, they still need to be even more specific earlier on about who the community is that they are addressing, namely the community of computational data scientists working in academia. This is mentioned later in the paper (p. 4), but ideally this information could be worked into the title and definitely it should be mentioned in the abstract.	We have added the following to the end of the abstract to clarify the audience: Other papers have focussed on many general rules for sharing data. This paper focusses on common sense ideas for sharing tabular data, for a target audience of academics working in data science adjacent fields who are about to submit for publication.
R1-2	The authors need to further clarify how their recommendations align with - and differ from - the slew of other recommendations for data management which exist (in articles - e.g https://doi.org/10.1371/journal.pcbi.1003542, and at the level of institutional/regional data stewardship - e.g. from the Research Data Alliance, or from libraries: https://www.monash.edu/library/researchers/n/researchdata/guidelines/sharing.) I think the key to differentiating this work is again making the community/audience explicit from the very beginning	We have added the following into section two, to help further describe our audience, and to describe how our paper is different to others:"We believe a key audience could be better served by the current literature on data sharing: researchers who have finished data analysis on tabular data, are ready to submit for publication. Many papers provide rules and guidelines for sharing data, such as providing metadata, citation, licenses, and lists of data repositories to use. These papers provide general advice from the ethical considerations of sharing data [@Moles2013], how to share many different data types (images, video, etc) [@Hart2016], to "Reward Colleagues Who Share Their Data Properly" [@Goodman2014]. There is also specific advice like the importance of very clean and properly shaped data [@White2013], and to use open source software with changes tracked [@Boland2017]. We walk a fine line between providing highly specific advice (e.g. how to specify NULL values) and the overly general (e.g. sharing data is good). Our goal is to provide practical advice for the researchers who are familiar with modern data science tools and develop their work like research compendia.
R1-3	Some of the recommendations themselves border on being too generic - e.g. 'use metadata with your data' - without specifying how metadata schemas/vocabularies can be chosen or what type of metadata are being considered. The examples that are provided (e.g. a list of variable names) are something that is perhaps more akin to including a codebook rather than structured metadata following a particular schema. Perhaps this overlaps with data documentation? Do the authors want to address more structured forms of metadata schemas which are often required by repositories?	We have added the following into the metadata section: A simple metadata file that describes the variable names, variable description, and unit text is a great starting point, and is sometimes known as a codebook. Metadata can be encoded into more formal plain text formats that follow a set standard, such as EML or SCHEMA-LD. A deep understanding of these formats isn't required to use them. For more details, see [@Broman2017; @Ellis2017; @Arslan2019 @Leipzig2021].

(continued)

ID	comment	response
R1-4	Along the same lines, it is unclear why the authors recommend using Zenodo and Dryad specifically to share data, aside from issues of data size (and I am not completely convinced by this argument either). Usually recommendations point researchers toward lists of repositories, searchable by subject, (e.g. re3data.org) which can be used to select an appropriate repository. Is there a reason for not doing this?	While we agree with the reviewer in general, R3Data is merely a long list of repositories. It does not provide curation or advice on best practices to make the data reusable. In contrast, Dryad and Zenodo are both general purpose non-commercial repositories that are leading the way in best practices for data and software deposition. So we have revised this section to note that researchers should follow community advice on domain repositories when available. Barring those, then Dryad and Zenodo are best suited for the purposes of reproducible data science. Revised text reads: "Data should be deposited into a relevant domain repository that generates an accession number such as a DOI. For example, gbif for species data, or genbank for genetics data. A exhaustive searchable list is available at re3data (https://www.re3data.org/) but provides no guidance to researchers. Data that do not fit appropriate domain repositories can be submitted to research data repositories **Zenodo** and **Dryad** which are currently leading the way in best practices for software and data. In addition to minting DOIs, they also provide citation templates."
R1-5	The structure of the paper could be improved. The article consists of two main parts: 1) the Introduction (which is really a section discussing the current state of incentivising data code and sharing for academic data scientists) and 2) Recommendations for data sharing. I recommend putting all of the recommendations and the current section 1.1 as subpoints in a second section devoted only to recommendations.	We have now changed this so the Introduction is section 1, and then the recommendations are section 2, with ssubsections 2.1 to 2.9
R1-6	The title of the paper does not seem to match the main gist of the paper. It takes some work for the reader to figure out that the recommendations are tied to incentivising data sharing. How exactly do the recommendations provide incentives for academic data scientists? Perhaps consider changing the title and/or spelling out the link between the two sections a bit more clearly.	We have altered the abstract to better capture out intended message. The title has...
R1-7	Perhaps it would help to spell out who the 'you' is that is being addressed in the recommendations?	This has been addressed by clarifying the audience, which has been done in the abstract and introduction

(continued)

ID	comment	response
R1-8	The treatment of data citations on p. 4 overlooks many issues about data citation and efforts which are underway. The authors state that 'datasets currently do not accrue citation credit since they are not tracked.' Not only are citations not tracked - many researchers don't actually cite data in the first place.	We have cited the most current work on data citations, but the rest of the reviwer comment restates what we have in the paper already.
R1-9	typo p.2 - around footnote #8 - there seem to be some missing phrases here?	Removed the unneeded 'and', changed, "In a study of applied computational research, and colleagues were only able to install" to "In a study of applied computational research, less than 20% of software described in publications was able to be installed and run."
R1-10	p.2 - What is meant by "trivial reproducibility?'	Removed 'trivial' - sentence changes from, "The barriers to verification are now low enough that trivial reproducibility is sometimes a click away" to "The barriers to verification are now low enough that reproducibility is sometimes a click away"
R1-11	p..6 - 'We don't have recommendations for non-tabular data.' The authors probably don't need to mention this, as this has already been made clear.	we can remove this sentence: https://github.com/karthik/ddd/blob/master/new-paper.Rmd#L130
R1-12	p. 9 - Reference 25 - Does the Piwowar paper really provide 'anecdotal' evidence?	NA
R1-13	A sentence in the abstract currently states "research data are only available to original investigators." A more accurate statement would be ""research data are often only available to original investigators."	Sentence changed from, "Despite the general appreciation of the benefits of data sharing, research data are available only to the original investigators" to "Despite the general appreciation of the benefits of data sharing, research data are often only available to original investigators."
R2-1	There are many other similar papers that have already been published on this topic, and I do not see that this work adds substantial new knowledge or advice. For instance: 10 Simple Rules for the Care and Feeding of Scientific Data (https://doi.org/10.1371/journal.pcbi.1003542), Nine simple ways to make it easier to (re)use your data (https://ojs.library.queensu.ca/index.php/IEE/issue/view/478), Ten Simple Rules for Digital Data Storage (https://doi.org/10.1371/journal.pcbi.1005097), Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing (https://doi.org/10.1371/journal.pcbi.1005278), and The Turing Way: A Handbook for Reproducible Data Science(http://doi.org/10.5281/zenodo.3233853).	Addressed in earlier comment

(continued)

ID	comment	response
R2-2	The title of the paper makes it sound like this will be a review of current incentives & barriers to data sharing and maybe some original research into this, but that is not what this paper is. This paper has an introduction that includes some information about how data isn't being shared, and then tips for what you should do when you share your data. As stated above, there are many other papers that share these same tips for sharing data, so in my opinion, this paper would be much stronger if it was indeed a more in depth discussion/research on what incentives are currently available and how they are working/not working.	We have changed the paper title and added information in the introduction to better articulate the audience we are addressing and that we think would benefit the most from reading it, and how it is different to the many other guides out there. While we believe that a paper which, provides "a more in depth discussion/research on what incentives are currently available and how they are working/not working" - our paper has a different goal - to describe the misalignment of these incentives, and to provide researchers with common sense approaches to sharing data.
R2-3	Section 9 is confusing. The authors say we can share data within a software package, but then spend a lot of time detailing why we shouldn't do that. What is the recommendation here?	We have added the following sentence to address this comment: " Nevertheless, the frictionless approach to sharing data in software packages makes reuse substantially faster than retrieving and correctly parsing files from a repository. We recommend the software approach only as an additive to proper data curation."
R2-4	The audience is a bit unclear to me. There is a mix of advice that is very general with little detail and then advice that is very specific. Who is the target audience for this advice?	we have clarified the audience in an earlier comment.
R2-5	The entire manuscript could benefit from more proof reading. There are several missing words, especially missing article/resource names - perhaps the citation manager removed the text of several citations? Also, the introduction is very long and has some issues with logical flow (e.g. the second paragraph switches context between progress and problems with reproducibility without any transitions). This could likely be resolved with another proof reading.	NA
R2-6	Figure 1 could use more explanation. The authors should introduce Nosek, and perhaps add a citation.	We have updated the citation.
R2-7	"There are numerous examples of similar problems in other fields, such as psychology, biology, and bioinformatics." Please add citations for this.	We have updated the citation.
R2-8	Watch for writing style that might discourage readers from implementing these ideas - for instance, remove phrases that are passive aggressive or off putting like "Just archive the data."	We have read through the paper and removed the phrase the reviewer mentioned, as well as checked the paper for clarity