Rigor

Background

This category covers situations where a dataset, published at a repository, stand alone or related to a research object (e.g. journal publication, preprint or other) is found to contain errors or gaps that call into question its validity for scientific use. The flaws may be related to: unintentional error, incomplete or partially available datasets, data manipulation or fabrication.

Examples

Cases may arise around human or tool error as well as situations where further refinement of experimental techniques or methods surfaces an issue related to published datasets. More specifically:

- Unintentional error e.g., errors in data collection, presentation (copy & paste errors, errors introduced by a tool e.g. Excel issue which reformatted numbers), calculation errors, refinement of experimental techniques or methods surfaces an issue in relation to a previously published dataset
- The dataset is incomplete or only partially available
- The dataset appears complete but it is not interpretable, due to ambiguous metadata descriptors
- Data manipulation or fabrication

How cases may arise

Concerns about a dataset may arise in the context of the dataset itself (record at data repository, data paper) or about a research object the dataset underlies (e.g. journal article, preprint, report).

If a concern arises in relation to potential flaws in a dataset, it is relevant to establish whether there are associated research outputs that may also need to be scrutinized. Concerns may be identified by different stakeholders at different stages of the data dissemination process:

- Author -- either after deposition of the dataset at a repository or after publication of a paper that makes use of the dataset
- Repository manager or data curator -- during the checks on datasets submitted for deposition at a data or institutional repository
- Editor or reviewer -- if reviewing the related dataset in conjunction to manuscript peer review
- Reader -- while utilizing the published data for their own research, or in the context of reading/assessing/reusing a journal article etc
- Institutional review board, ethics board, or other institutional research administrators -- in the context of an institutional investigation about published work or the author, or funding agency verifying compliance with a mandated data policy

In a first instance, it is recommended to raise any concerns directly with the author and the host of the dataset (e.g. data repository, journal), rather than via public commentary for example on social media or blogging sites.

Recommendations

In all the scenarios below, the data publisher which identified or which first received that first received the concern (e.g., data repository, journal) should take reasonable steps to establish whether another party (e.g., related journal) should be notified and where necessary, communicate to the other party that an issue has arisen. It may not always be possible for a data repository to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset. Once a resolution is reached the data publisher that first received the concern should notify this to the person raising the concerns.

What actions should be taken if the dataset has not yet been published? Who needs to be involved in this decision?

If the concern relates to a dataset that has been submitted but has not been yet published (e.g. it is undergoing checks at the data repository, the data repository provides functionality for private access to an unpublished dataset, dataset is under review at a journal), the issue may be identified by the authors themselves, by the data repository or by a reviewer or editor at a journal.

- Notified parties (e.g., repository) should follow up with the author noting the issue and asking for an updated dataset
- Repositories: If a revision is possible, the author should revise their dataset and/or its associated metadata. If the issues are too major and make a revision not viable, the author should withdraw the data deposition from the repository
    - If the author disputes the issue but the data publisher has remaining concerns, the data publisher may opt to not publish the data
    - If there are concerns about research practice/integrity - the data repository should consider contacting the author's institution to raise the concerns
    - If there is a paper associated with the dataset under consideration at a journal and the journal name is known to the repository, contact journal
- Journal publishers: Ask for authors to address through a revision. If the issues are too large to address via revision or impact the conclusions of the manuscript, the author should withdraw the manuscript from review.
    - If the extent of the impact on the paper is unclear or there are integrity concerns (e.g. related to potential data manipulation or fabrication), the journal may wish to seek input by an expert - if the expert confirms major issues, then the journal would reject the submission.
- Journals may also consult the COPE flowchart for handling concerns about data integrity in a submitted manuscript, available here: https://publicationethics.org/files/Fabricated%20data%20A.pdf

What actions should be taken for a published dataset? Who needs to be involved in this decision?

- Repositories:
    - Contact the corresponding author about the issue if they did not raise it themselves
    - Ask the author to submit a new version of the dataset if applicable
    - If the dataset cannot be modified (e.g., it's incomplete, data is lost, data is falsified), post a notification on the dataset that indicates the concerns around scientific rigor associated with the prior dataset version (see below)
    - If the author disputes the issue or does not update the dataset record, the data repository should post a notification alerting potential users to the issues with the dataset within a reasonable amount of time. There may be instances where rigor imposes a risk, the repository can take reasonable steps to assess if this is the case
    - If there are concerns about research practice/integrity (e.g. misconduct raised by a co-author involved with the dataset, documented concerns from a user/reader that the dataset is fabricated), consider contacting the corresponding author's institution to raise the concerns. This is expected to be a step that only applies in

rare, exceptional situations where there are major concerns about data manipulation, fabrication or falsification and the data publisher has received no response or an inadequate response from the author(s)
    - If there is a clear relation to a report or published paper, contact relevant publisher
- Journal publishers:
    - The journal may seek input by an expert, if required, to establish how the issue with the dataset impacts the article
    - If a new dataset version addresses the issue and there is a new dataset DOI, add a correction to the published paper pointing users to the latest dataset version
    - If the issues cannot be resolved via a new version of the dataset, or if there are concerns about research practice/integrity, the journal should review how the issues impact the published article and whether an Expression of Concern or a retraction should be issued
- Journals may also consult the COPE flowchart for handling concerns about data integrity in a published article, available here: https://publicationethics.org/files/Fabricated%20data%20B.pdf
- Note that there may be more than one output associated with a single dataset, and that the research output may or may not be impacted depending on the concerns about the dataset. The host of the research output (e.g. journal) would need to establish whether action is required on the output they host. In addition, a single research output may rely on more than one dataset so it would also be relevant for the host (e.g. journal) to follow up to establish how the output is impacted depending on whether or not individual datasets have concerns about their rigor or completeness.

To whom and when does it need to be reported?

The recommendation is for the corresponding author to take the lead in communicating issues to relevant parties in relation to the relevant copies of the dataset they host (i.e. communicate to data repository, journal etc).

In situations where the author is not responsive or provides an unsatisfactory response, and major concerns remain or have been established about the dataset, the recommendation is that the organization that is informed of the issue will take reasonable steps to inform other parties that host a public record of the dataset or a research object associated with it. This means that a data repository should notify a journal if there are associated articles related to the dataset (when information is available), and vice versa for applicable situations.

How should the public be notified?

- Repositories:
    - Update to the data record with updated metadata outlining errors in the original version and updates in the new version
    - When an update is not possible, post a public notification on the dataset outlining confirmed flaws or community concerns, and warning the readers about future use of the dataset to build on the research
        - If there is information available elsewhere outlining issues about the dataset (e.g. in a journal Expression of Concern or retraction), the notification may refer to that for further context
        - Note that the future reuse of the dataset as a designated control flawed dataset may be suitable for certain types of studies

- Journal publishers:
    - Journals may post interim comments on the article (if the feature is available) if they wish to alert readers to a concern about the data while their follow up is ongoing
    - Correction to the article record via a Correction, Expression of Concern or Retraction, according to the extent to which the issues impact the standing of the published work
    - The journal may consider a public statement if the concerns involve high-profile publications or a large group of articles

How do we handle inaction or silence from stakeholders (e.g, the publisher, the authors, the institution)?

The below outline steps to take in situations where the party which initiates action does not receive a response from the other party, of where the other party fails to take action or to do so in a timely manner - we acknowledge that workflows vary from one organization to another and that the extent and length of the process required to handle issues may vary.

The recommendation is for the corresponding authors to take the lead in notifying relevant parties of issues with their dataset. If the corresponding author of the dataset/the corresponding author for the article is unresponsive when contacted by the data repository/journal, a follow-up communication should be attempted copying all authors. If no response is received from any of the authors, the parties hosting the dataset or associated research output may issue public notifications outlining the concerns with the dataset, and if they deem it necessary raise the concerns to the attention of the authors' institution to investigate.

If the concern is raised to a journal and this fails to respond, the issue can be raised to the attention of their publisher and, failing that, to COPE for review if the journal is a COPE member. If a data repository is involved, this may choose to issue a notification on their records, according to their frameworks, without the input of the journal. In the lack of response by the journal/publisher, the matter may be raised to the funding body for the research and/or the author's institution.

If the institution fails to respond, the issue can be raised to the attention of a national regulatory body or the author's funder. The data repository and/or journal may choose to issue a notification on their records, according to their frameworks, without the input of the institution.

Resources

- Discrepancy between data and article findings: https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-015-0847-3 -
- Potential contamination and data error: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1802-z
- Errors in data classification: https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/1471-2393-14-202
- Author identified errors in raw data: https://www.sciencedirect.com/science/article/pii/S0195666309000038?via%3Dihub
- Concerns over data duplication and potential manipulation: https://royalsocietypublishing.org/doi/10.1098/rspb.2020.0077
- Retracted datasets at repository due to errors in calculation and analysis in accompanying article: https://springernature.figshare.com/articles/dataset/RETRACTED_DATASET_Paired_watershed_study_data_and_related_statistical_model_predictions_to_investigate_the_impact_of_forest_removal_and_planting_on_water_yield/7770035