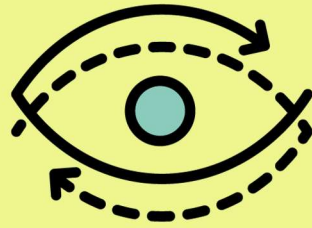


REPORT



TRUSTparenity

D1.1 A collection of existing assessed and categorized interventions

Authors: Panagiotis Kavouras, Pooja Aryal, Ilse de Jong, Joeri Tjndink

Reviewers: Vivian Nchanchou Mbanya, Rosemarie de La Cruz Bernabe

Project title: Increasing reproducibility through the co-creation of interventions that support a transparent and trustworthy research ecosystem: TRUSTparenity

Project acronym: TRUSTparenity

Grant Agreement no.: 101188172

Lead contractor for this deliverable: University of Oslo

Deliverable factsheet:

Project Number:	101188172
Project Acronym:	TRUSTparenity
Project Title:	Increasing reproducibility through the co-creation of interventions that support a transparent and trustworthy research ecosystem: TRUSTparenity
Title of Deliverable:	A collection of existing assessed and categorised interventions
Work Package:	WP1
Due date of deliverable	M13 – 28/2/2026
Actual date of submission	28/2/2026
Authors:	Panagiotis Kavouras, Pooja Aryal, Ilse de Jong, Joeri Tjldink
Contributors:	Signe Bovard Urup, Karen Kjølby Eika, Daniele Fanelli, Viktor Hinzl, Vana Stavridi, Dimitrios Koulos, Leonidas Ananiadis, Eleni Spyrou, Mariska M.G. Leeflang, Mads P. Sørensen
Reviewers:	Vivian Nchanchou Mbanya, Rosemarie de La Cruz Bernabe

ABSTRACT:	This report presents the results of the scoping review and of the Delphi survey conducted during the first year of TRUSTparenity, in the context of Tasks 1.1 and 1.2. The main aim of the scoping review was to collect interventions that promote reproducibility. The methodology of the scoping review was based on the one jointly conducted by TIER2 and OSIRIS Research and Innovation Actions; the report highlights the points of departure from that methodology and the reasons for this. The scoping review covered the period from August 2023 (the last month the TIER2 and OSIRIS scoping had covered) until March 2025; 68 publications that fulfilled the inclusion criteria were found, from which the process of data extraction was carried out. Here we report the findings of the scoping review, i.e. reproducibility interventions, across a number of parameters, like – among others – disciplinary focus, placement at the research cycle, and potential implementers of the reproducibility interventions. The main aim of the Delphi survey was to create an expert-informed taxonomy of types of interventions for facilitating reproducible research practices. The taxonomy developed is based on the different stages of the research cycle plus one category relevant to infrastructures. The results of the scoping review and the Delphi survey were combined by placing the interventions found by this scoping review, plus those found by the TIER2/OSIRIS projects, at the different taxonomic categories. The combination of the scoping review and Delphi survey provides a “seed” for the project’s Toolbox of reproducibility interventions, with the scoping review providing the content of it and the Delphi survey providing the structure of it.
Keyword List:	Reproducibility interventions, Classification, Categorisation, Scoping review, Delphi survey.

Consortium

	Role	Name	Short Name	Country
1.	Coordinator	University of Oslo	UiO	NO
2.	Partner	Aarhus University	AU	DK
3.	Partner	Science of Science Ltd.	SoS	GB
4.	Partner	Amsterdam UMC	AUMC	NL
5.	Partner	University of Latvia	UoL	LV
6.	Partner	National Technical University of Athens	NTUA	EL
7.	Partner	Finnish National Board on Research Integrity	TENK	FI
8.	Partner	Aristotle University of Thessaloniki	AuTh	EL
9.	Partner	University of Santa Catarina	UFSC	BR
10.	Partner	National University	NU	PH
11.	Partner	Kwame Nkrumah University of Science and Technology	KNUST	GH
12.	Partner	Lviv Polytechnic National University	LPNU	UA
13.	Associated Partner	Open Science NL	OSNL	NL

Revision history

Version	Date	Revised by	Reason
0.1	06.02.2026	Panagiotis Kavouras	First draft of scoping review report
0.2	16.02.2026	Pooja Aryal, Ilse de Jong	Comments on the first draft of scoping review report
0.3	20.02.2026	Ilse de Jong	First draft of the Delphi survey report
0.4	23.02.2026	Panagiotis Kavouras, Joeri Tijding	Comments on the first draft of the Delphi survey report
0.5	25.02.2026	Panagiotis Kavouras	Compilation of the prefinal vesion
0.6	26.02.2026	Pooja Aryal, Vivian Nchanchou Mbanya, Ilse de Jong, Joeri Tijding, Rosemarie de La Cruz Bernabe	Comments on the prefinal version
1.0	27.02.2026	Panagiotis Kavouras	Compilation of the final vesion

Table of contents

1. Executive summary.....	5
2. Scoping review	5
2.1 Objective	5
2.2 Methodology	6
2.3 Results.....	9
3. Delphi survey	17
3.1 Objective.....	17
3.2 Methodology	17
3.3 Results.....	18
3.4 Next steps	28
4. References.....	28
5. Annexes.....	29
5.1 Search strategies	29
5.2 Inclusion and exclusion criteria	33
5.3 Articles for data extraction – bibliographic records in alphabetical order	35
5.4 Data extraction items and their answering options.....	41
5.5 Data extracted (1/3) – Context of the intervention	44
5.6 Data extracted (2/3) – Characteristics of the intervention	48
5.7 Data extracted (3/3) – declared limitations of the study.....	59
5.8 Delphi round 1 survey	69
5.9 Feedback report for round 1.....	76
5.10 Delphi round 2 survey	99
5.11 Reference numbers to the combined results of the TIER2/OSIRIS and TRUSTparenacy scoping reviews	108

1. Executive summary

This report presents the results of the scoping review and of the Delphi survey conducted during the first year of TRUSTparenacy, in the context of Tasks 1.1 and 1.2, respectively. The methodology of the scoping review was based on the one jointly conducted by TIER2 and OSIRIS Research and Innovation Actions (RIA); the report highlights the points of departure from that methodology and the reasons for this. The initial automatic screening of titles was conducted by the system developed by iRISE RIA. This heavy reliance on the experience and outputs created by all three »mother« reproducibility-related projects was instrumental for TRUSTparenacy to meet the targets of Tasks 1.1 and 1.2 effectively and efficiently.

The main aim of the scoping review was to collect interventions that promote reproducibility. The scoping review covered the period from August 2023 (the last month the TIER2 and OSIRIS scoping had covered) until March 2025; after the implementation of four screening stages (automatic, title/abstract, full text, and deduplication with the results of the TIER2/OSIRIS results) 83 publications that fulfilled the inclusion criteria went through the data extraction process. In the end the reported extracted data were obtained from 68 publications, since 15 additional were excluded during this stage (Figure 1). Here, we report the findings of the scoping review, across 9 parameters, e.g. disciplinary focus, potential implementers, and placement at the research cycle of the reproducibility interventions. The reproducibility interventions found are, predominantly, relevant to Medical and Health Sciences, are, predominantly, implemented by researchers, and they are, predominantly, implemented during data gathering.

The main aim of the Delphi survey was to create an expert-informed taxonomy of types of interventions for facilitating reproducible research practices. The taxonomy developed is structured across five categories, four based on different stages of the research cycle (Preparation, Data collection, Data analysis, and Dissemination) plus one relevant to infrastructures. These five main categories are broken down to 17, in total, sub-categories (Figure 11).

The results of the scoping review and the Delphi survey were combined, by the leaders of Tasks 1.1 and 1.2, UiO and AUmc respectively, by placing the interventions found by this scoping review, plus those found by the TIER2/OSIRIS projects (173 in total), at the different categories and sub-categories that emanated from the Delphi survey (Table 16). The combination of the scoping review and Delphi survey provides a »seed« for the project's Toolbox of reproducibility interventions, to be developed by Task 3.2, with the scoping review providing the content of it and the Delphi survey providing the structure of it. The project's Toolbox of interventions will provide practical guidance not only for researchers within these fields, but also for institutions, publishers, funders, and other relevant stakeholders seeking to effectively support reproducible research.

2. Scoping review

2.1 Objective

The objective of this scoping review is to extend, timewise, the one that was conducted in the context of the TIER2 and OSIRIS projects. The TIER2/OSIRIS scoping review captured publications until August 2023; this scoping review extended it until the end of March 2025. We use the term »extension« to highlight the fact that we utilised the same methodology, i.e. common understanding of what comprises a reproducibility practice, same search strategies, same databases, among others; details are presented

at the methodology section. The aim, similarly to that of the TIER2/OSIRIS scoping review was to enrich the existing overview of interventions that aim to improve reproducibility and replicability in scientific research.

2.2 Methodology

A Scoping Review was used to perform evidence synthesis on heterogeneous areas. The heterogeneity comes from the fact that (a) the conceptualisation of reproducibility and replicability and the terminology used varies across disciplines and (b) the relevant literature comes from a wide variety of research fields/disciplines. The Scoping Review was conducted following the guidance of the PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation (Tricco et al., 2018).

2.2.1 Search strategy

The search strategy was obtained from the TIER2/OSIRIS scoping review (see Annex 5.1), as well as the databases used, namely Medline, Embase, Web of Science, PsycINFO, Scopus, and ERIC. Despite the fact that the TIER2/OSIRIS scoping review included August 2023 as the last month, we included August 2023 as the first month of our Scoping Review for two reasons: (a) depending on the database used, the TIER2/OSIRIS scoping review included different ending dates within August 2023, so there might be some studies that had been published at the end of August and, as a result, would not have been included in the results and (b) we aimed at checking whether there would be duplicate results within August 2023; in case the TIER2/OSIRIS scoping review had identified relevant publications within August 2023, our scoping review should have also identified them. If not, we were going to try to interpret why this was not the case. This could be conceived, in very loose terms, as a very focused – timewise – *conceptual* replication exercise.

The *conceptual* term needs explanation, since it could be argued that for August 2023 this would be a *direct* and not a conceptual replication exercise. The term conceptual was used above, because before the initiation of the title/abstract screening process there was a need for an additional step that was not part of the TIER2/OSIRIS scoping review. Specifically, the search that was performed in February 2025 (spanning from August 2023 until the – then - current date), resulted in 16907 publications. This was significantly higher than expected, based on the fact that our Scoping Review was covering roughly two years. It was collectively decided that screening such a high number of publications would result in a delay, taking into account the tight workplan. Specifically, the results should be reported internally to the Task 1.2 partners no later than the end of summer 2025. This was considered not feasible and an alternative strategy was followed.

We relied on the studies that were identified by the iRISE project, whose “living” database of studies (a database that was regularly updated) was last updated on 16 March 2025, i.e. it included roughly one more month with regard to our database of 16907 publications. The iRISE database included 20792 publications that, before sent to us, had been automatically screened for relevance using iRISE's machine learning classifier. This performed, according to our colleagues from iRISE, a screening with a sensitivity of 95% and a specificity of 77%, meaning it is over-inclusive, so the risk of not identifying potential relevant publications was low. Eventually, iRISE provided to us the screened database that included 4750 articles between 2023 and March 2025. That was the cohort of publications on which we initiated the title/abstract screening process.

Inclusion/exclusion criteria

The inclusion/exclusion criteria we used were based on those developed for the TIER2/OSIRIS scoping review and they are reported in Annex 5.2. The most challenging task was to create a common

understanding, among the team that conducted the title/abstract and full text screening, of what comprised a valid test of a reproducibility intervention. Since the interventions that were going to be selected by the scoping review are going to be brought to the attention of end users as part of institutional guidelines that promote reproducibility, the Reproducibility Promotion Plans, we had to apply relatively strict criteria. In the end, the studies we perceived to provide evidence for the effectiveness of a reproducibility intervention were greatly limited, with regard to the studies retrieved by the search strategy.

The inclusion/exclusion criteria that were used by the TIER2/OSIRIS projects were modified, with the aim of making as clear as possible that data extraction was going to be applied only to studies that had tested the intervention on whether it had any effect on reproducibility. The need to modify the TIER2/OSIRIS inclusion/exclusion criteria was imposed from the fact that at the start of the title/abstract screening there was a high percentage of conflicts. After collectively modifying the inclusion/exclusion criteria a smaller team of four members of the screening group ran a pilot screening exercise of 20 randomly selected studies from those that had been identified. The result of the exercise was that there was consensus on the include/exclude decision on two thirds of the randomly selected cohort of studies. The other one third that produced conflict was examined by all four members and a consensus on the include/exclude decision was achieved for all but four studies. This was considered a satisfactory outcome of the exercise; the success was based on the modification of the inclusion/exclusion criteria. In Annex 5.2, we present the original inclusion/exclusion criteria of the TIER2/OSIRIS scoping review and the modified ones in tabular form that facilitates comparison.

2.2.2 Screening process

Title and abstract screening was performed within 24/03/2025 and 28/04/2025 by (in alphabetical order): Leonidas Ananiadis (NTUA), Pooja Aryal (UiO), Rosemarie Bernabe (UiO), Karen Kjølby Eika (AU), Daniele Fanelli (SoS), Viktor Hinzl (AU), Ilse de Jong (AUMc), Panagiotis Kavouras (UiO), Dimitrios Koulos (NTUA), Mariska M.G. Leeflang (AUMc), Mads P. Sørensen (AU), Eleni Spyarakou (NTUA), Vana Stavridi (NTUA), Joeri Tjldink (AUMc), and Signe Bovard Urup (AU). The conflicts were resolved by Pooja Aryal and Panagiotis Kavouras within 08/03/2025 and 30/04/2025.

Full text screening was performed within 01/05/2025 and 30/06/2025 by Pooja Aryal (UiO), Rosemarie Bernabe (UiO), Daniele Fanelli (SoS), Panagiotis Kavouras (UiO), Mariska M.G. Leeflang (AUMc), Eleni Spyarakou (NTUA), Vana Stavridi (NTUA), and Joeri Tjldink (AUMc). The conflicts were resolved by Pooja Aryal and Panagiotis Kavouras within 15/05/2025 and 01/07/2025.

As mentioned above, the 20792 publications that were retrieved from the iRISE database, 4750 publications were identified for the initiation of the screening process, after the automatic screening performed by iRISE (as described above in detail). Of the 4750 publications, 54 in total were found to be duplicates; 33 duplicates were automatically detected by Covidence and 21 we detected manually, during the title and abstract screening. Figure 1 depicts the PRISMA diagram of the screening process.

Based on title and abstract screening, 4363 publications were excluded, resulting in 333 publications for the full text screening; the full texts of all publications were retrieved and went through full text screening. Of the 333 publications, 244 were excluded for various reasons, listed in Figure 1. Of the 89 studies selected for data extraction, 6 were removed, based on the deduplication we conducted with the studies found from the TIER2/OSIRIS scoping review. This resulted in a total of 83 publications for data extraction.

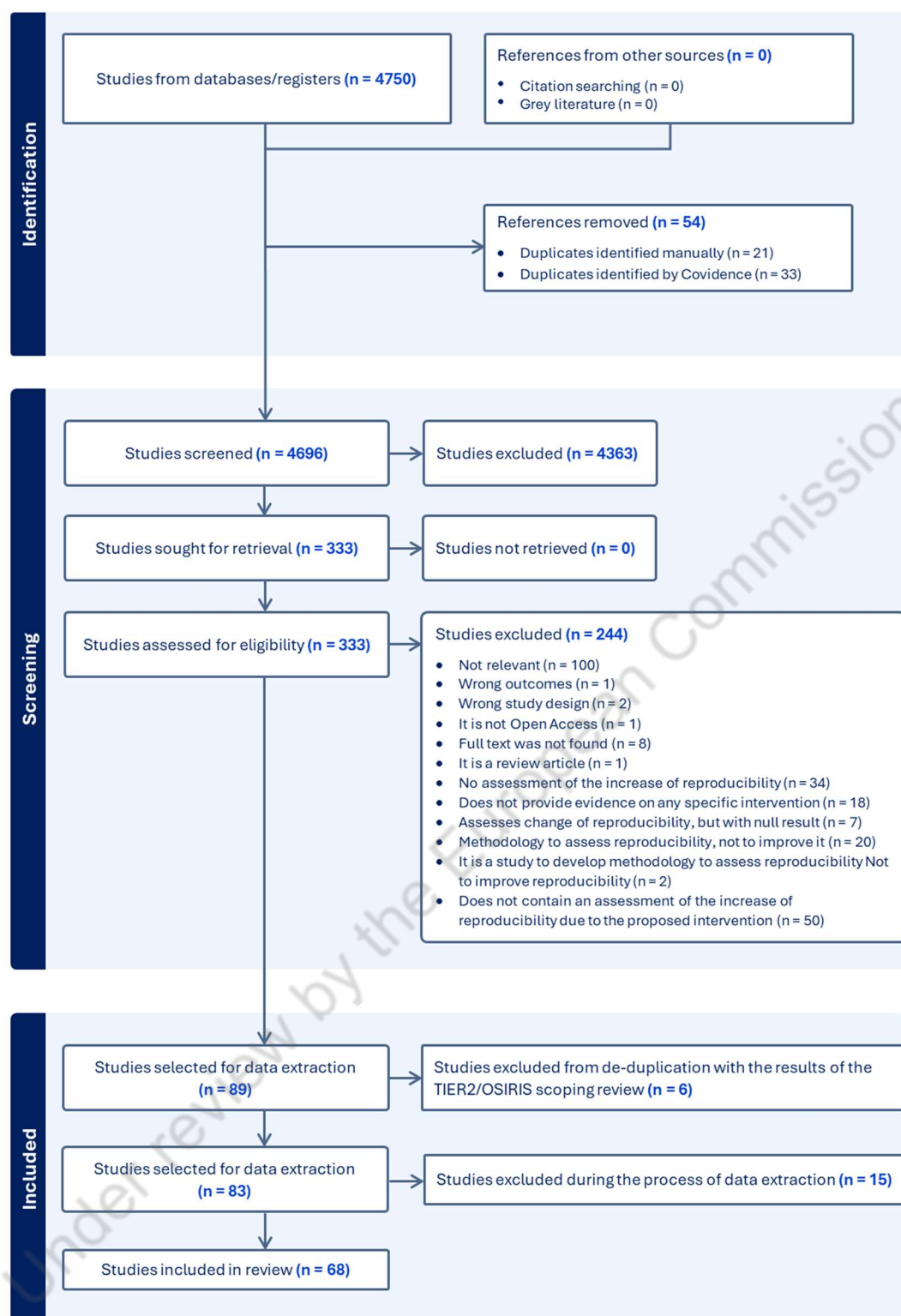


Figure 1: PRISMA flowchart of the screening process.

2.2.3 Data extraction

Despite the fact that these 83 publications had already gone through three screening stages (automatic, title/abstract, and full text), during data extraction, 15 additional publications were excluded as non-relevant. These were excluded since they either did not contain evidence for testing the interventions or they were not about reproducibility whatsoever, despite the fact the increase of reproducibility was

mentioned as a possible outcome. The results presented below came from the data extraction process from a total of 68 publications; their bibliographic record is presented in Annex 5.3.

Data extraction was performed by Pooja Aryal and Panagiotis Kavouras, of UiO, from 28/07/2025 to 21/08/2025. The group of data extraction items we used was a subset of those used for the TIER2/OSIRIS scoping review. After consulting the lead researcher of the TIER2/OSIRIS scoping review, Mariska M.G. Leeflang (AUMc), who is also a member of the TRUSTparenacy consortium, we simplified, in some cases, the list of inclusion/exclusion criteria, for reasons of effectiveness and efficiency. The data extraction items, together with the answering options, are listed in Annex 5.4. At the same Annex we have listed, in a separate table, the TIER2/OSIRIS scoping review data extraction items that were excluded from our scoping review.

All extracted data are presented in three annexes: Annex 5.5 presents the extracted data relevant to the context of the intervention; Annex 5.6 presents the extracted data relevant to the characteristics of the interventions; Annex 5.7 presents the extracted data relevant to the declared by the authors limitations of the studies.

2.3 Results

Here, we present the results of the scoping review, across a number of parameters that represent some of the extraction data items. The parameters have been selected with a view to the descriptors that will be needed for the toolbox of interventions that is being developed in the context of Task 3.2., i.e. some of these parameters will be considered to describe the interventions found in the publications – among other parameters – that will be included in the project's Toolbox.

2.3.1 Publications per academic field

Table 1 lists the number of publications per academic field and the respective reference number of the publication, taken from Annex 5.3.

Table 1: Number of publications per academic field.

Academic field	Number	References
Agricultural and veterinary sciences	3	15, 16, 28
Engineering and technology	9	2, 10, 17, 32, 33, 47, 61, 65, 66
Humanities and the arts	2	42, 51
Medical and health sciences	34	1, 3, 4, 5, 6, 7, 8, 9, 14, 19, 21, 22, 23, 24, 25, 26, 30, 35, 37, 38, 39, 43, 49, 50, 52, 53, 55, 56, 57, 59, 60, 64, 67
Natural sciences	10	11, 13, 27, 29, 41, 44, 46, 48, 63, 68
Social sciences	6	12, 18, 31, 34, 36, 54
Not applicable/Not reported	4	40, 45, 58, 62

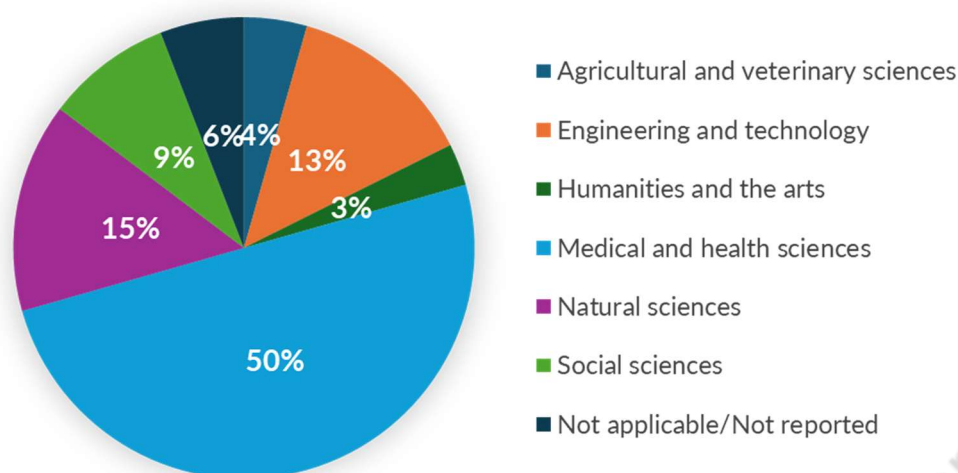


Figure 2: Allocation of interventions per discipline.

Figure 2 is a pie chart of Table 1, where quantitative information can be more easily visualized. Most of the publications, half of them, are relevant to the field of Medical and Health sciences. Natural sciences and Engineering and technology are each contributing roughly one sixth of the publications.

2.3.2 Publications per implementer

The table below lists the number of publications with regard to the response to the question “Who implements the intervention?” and the respective reference number of the publication, taken from Annex 5.3.

Table 2: Number of publications with regard to the response to the question “Who implements the intervention?”

Who implements the intervention?	Number	References
Academic staff	2	7, 24
Journal	1	8
PhD researchers	1	1
Researchers	64	2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68

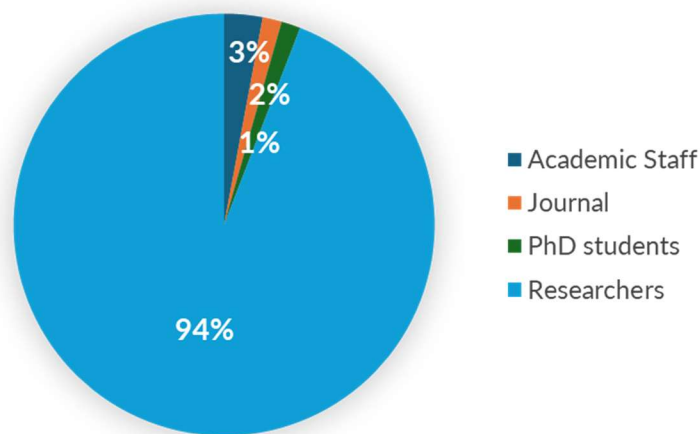


Figure 3: Allocation of interventions per relevant type of stakeholder that implements the intervention.

Figure 3 is a pie chart of Table 2, where quantitative information can be more easily visualized. The vast majority of publications, have studied reproducibility interventions that are implemented by researchers.

2.3.3 Publications per affected type of stakeholder

The table below lists the number of publications with regard to the response to the question “Who is affected by the intervention?” and the respective reference number of the publication, taken from Annex 5.2.

Table 3: Number of publications with regard to the response to the question “Who is affected by the intervention?”

Who is affected by the intervention?	Number	References
Institution	1	21
Journal	1	8
PhD students	1	24
Researchers	65	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68

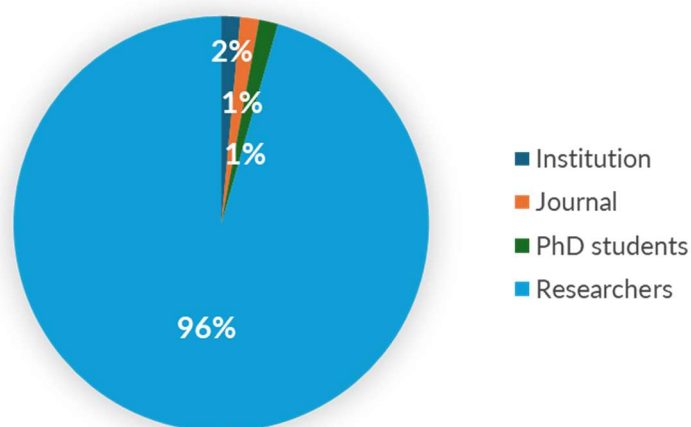


Figure 4: Allocation of interventions per affected type of stakeholder.

Figure 4 is a pie chart of Table 3, where quantitative information can be more easily visualized. The vast majority of publications, have studied reproducibility interventions that affect researchers.

2.3.4 Publications per research stage

The table below lists the number of publications with regard to placement of the intervention at a specific stage of the research process and the respective number of the publication, taken from Annex 5.3.

Table 4: number of publications with regard to placement of the intervention at a specific stage of the research process.

Stage of the research process	Number	References
Before research	4	23, 28, 31, 38
During research	48	1, 2, 4, 5, 7, 9, 10, 11, 13, 14, 15, 16, 17, 20, 21, 22, 25, 26, 29, 30, 32, 33, 35, 37, 39, 40, 41, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 57, 60, 61, 62, 63, 64, 65, 66, 67, 68
After research before or at publication	13	6, 8, 12, 18, 19, 27, 34, 36, 42, 43, 56, 58, 59
After publication	3	3, 24, 49

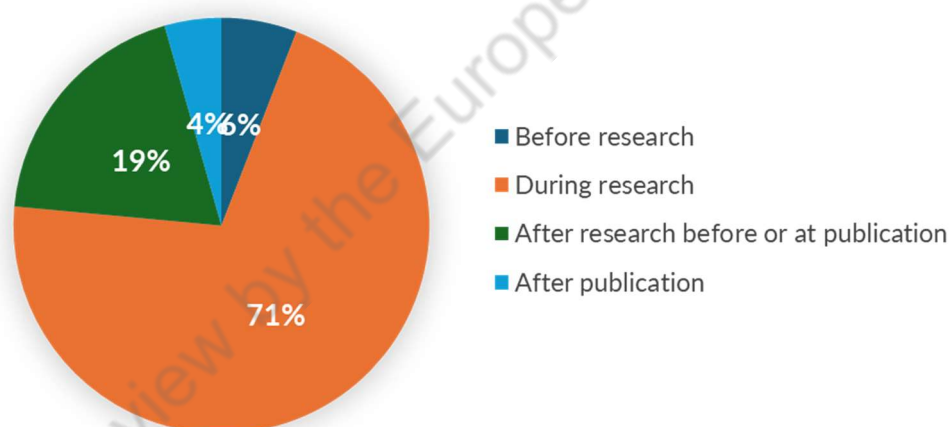


Figure 5: Allocation of interventions per research stage.

Figure 5 is a pie chart of Table 4, where quantitative information can be more easily visualized. The majority of publications have studied reproducibility interventions that are applied during the conduct of research. The second most significant contribution comes from reproducibility interventions that are applied after research and before or at publication.

2.3.5 Publications per intervention classification

The table below lists the number of publications with regard to the type of intervention, according to the classification of the TIER2/OSIRIS scoping review and the respective reference number of the publication, taken from Annex 5.3.

Table 5: Number of publications with regard to the type of intervention.

Intervention classification	Number	References
Open data and materials	9	3, 4, 8, 13, 19, 20, 27, 42, 43
Open methodology	21	1, 5, 6, 9, 10, 15, 16, 18, 21, 22, 29, 30, 31, 32, 35, 38, 45, 54, 65, 67
Open science policies and guidelines	25	12, 24, 34, 36, 40, 41, 44, 46, 47, 48, 49, 51, 52, 53, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 68
Open science tools	11	2, 7, 11, 14, 17, 23, 25, 26, 33, 37, 66
Other	2	28, 50

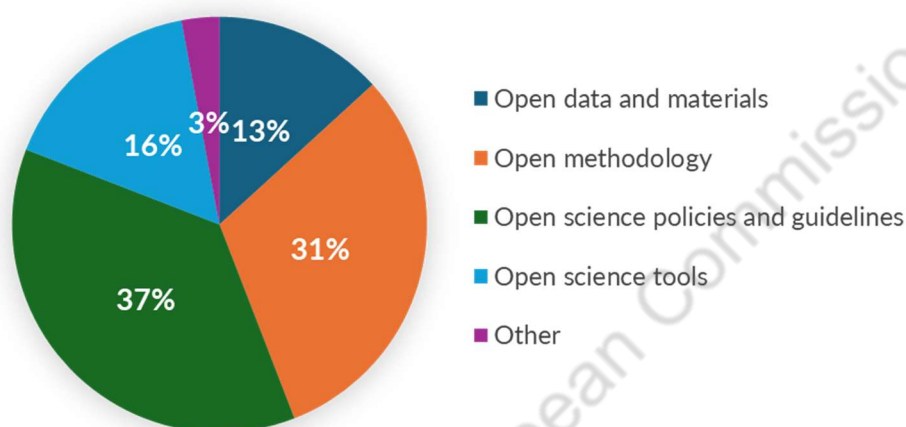


Figure 6: Allocation of interventions per classification.

Figure 6 is a pie chart of Table 5, where quantitative information can be more easily visualized. Two types of interventions were found to be more predominant, each one having been studied by roughly one third of the publications, namely interventions classified as “Open Science policies and guidelines” and “Open methodology”. In addition, two types of interventions, each one having been studied by roughly one sixth of the publications, were classified as “Open Science tools” and “Open data and materials”.

2.3.6 Publications per intervention sub-classification

The table below lists the number of publications with regard to the type of intervention, according to the sub-classification of the TIER2/OSIRIS scoping review and the respective reference number of the publication, taken from Annex 5.3.

Table 6: Number of publications with regard to the type of intervention.

Intervention sub-classification	Number	References
data availability	9	3, 4, 8, 13, 19, 20, 27, 42, 43
on data/code sharing or open science practices	11	1, 24, 34, 47, 49, 51, 53, 57, 59, 61, 62
on reporting checklists or guidance	4	40, 41, 52, 56
on research quality	4	44, 48, 55, 63
open protocol	15	5, 9, 10, 16, 18, 21, 22, 29, 30, 32, 35, 38, 39, 54, 67
open workflow tools, workflow, management systems	11	2, 7, 11, 14, 17, 23, 25, 26, 33, 37, 66

policy guidelines (e.g. of funders/publishers)	1	12
pre-registration, study registration, analysis plan	1	31
project workflow	4	6, 15, 45, 65
reporting guidelines, reporting standards	6	36, 46, 58, 60, 64, 68

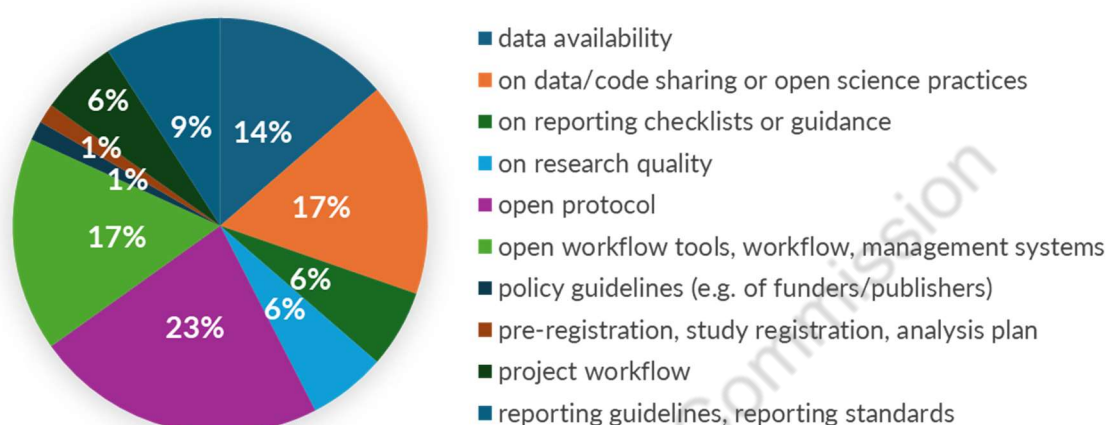


Figure 7: Allocation of interventions per sub-classification

Figure 7 is a pie chart of Table 6, where quantitative information can be more easily visualized. The type of intervention found to be more predominant, has been studied by roughly one fourth of the publications, namely intervention that belongs to the sub-classification of “Open protocol. In addition, three types of interventions, each one having been studied by roughly one sixth of the publications, belonged to the sub-classifications of “Open workflow tools, workflow, management systems”, “on data/code sharing or open science practices”, and “Data availability”.

2.3.7 Publications per type of outcome

The table below lists the number of publications with regard to the type of outcome they produce, according to the terminology of the TIER2/OSIRIS scoping review and the respective reference number of the publication, taken from Annex 5.3.

Table 7: Number of publications with regard to the type of outcome they produce

Intervention outcome	Number	References
Computational/Outcome/Result/Numerical/Statistical reproducibility	34	1, 7, 8, 9, 13, 20, 25, 28, 36, 40, 41, 42, 43, 44, 46, 47, 49, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 68
Inferential reproducibility	2	45, 50
Method /methodological reproducibility	31	2, 3, 4, 5, 6, 10, 11, 12, 14, 15, 16, 18, 19, 21, 22, 23, 24, 26, 27, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39, 48, 67
Other	1	17

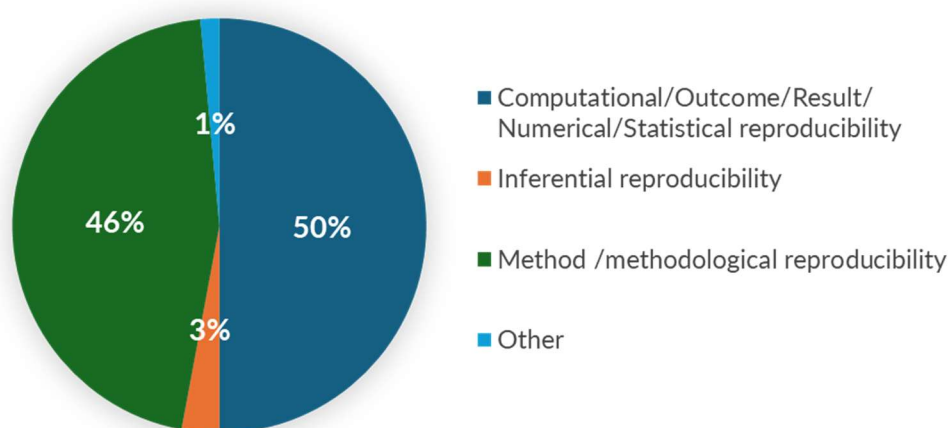


Figure 8: Allocation of interventions per type of outcome.

Figure 8 is a pie chart of Table 7, where quantitative information can be more easily visualized. Two were the most predominant types of outcomes of the interventions that were studied, each by almost half of the publications: “Computational/Outcome/Result/Numerical/Statistical reproducibility” and “Method /methodological reproducibility”.

2.3.8 Publications per type of proxy outcome class

The table below lists the number of publications with regard to the type of proxy outcome class they belong to, according to the terminology of the TIER2/OSIRIS scoping review and the respective reference number of the publication, taken from Annex 5.3.

Table 8: Number of publications with regard to the type of proxy outcome class they belong to.

Intervention proxy outcome class	Number	References
Data	17	8, 19, 20, 22, 24, 27, 43, 52, 54, 56, 57, 58, 59, 60, 64, 65, 66
Methods transparency	50	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 18, 21, 23, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 44, 45, 46, 47, 48, 49, 50, 51, 53, 55, 61, 62, 63, 67, 68
Research material sharing	1	17

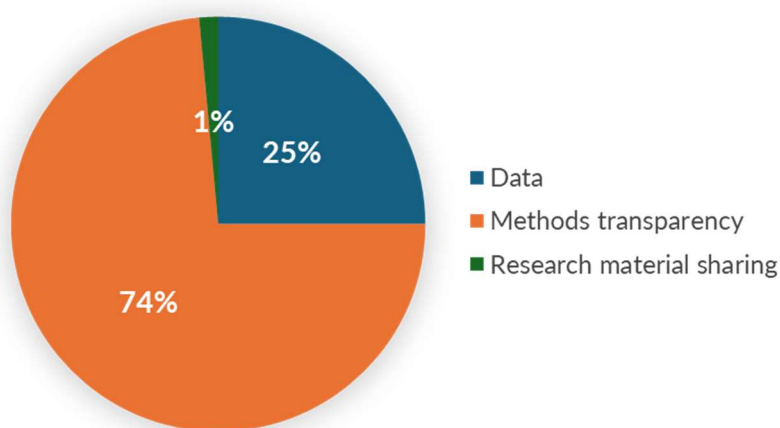


Figure 9: Allocation of interventions per type of proxy outcome class.

Figure 9 is a pie chart of Table 8, where quantitative information can be more easily visualized. Almost three thirds of the publications studied interventions whose proxy outcome class is “Methods transparency”. The other one fourth of publications studied interventions whose proxy outcome class is “Data”.

2.3.9 Publications per type of proxy outcome sub-class

The table below lists the number of publications with regard to the type of proxy outcome sub-class they belong to, according to the terminology of the TIER2/OSIRIS scoping review and the respective reference number of the publication, taken from Annex 5.3.

Table 9: number of publications with regard to the type of proxy outcome sub-class they belong to.

Intervention proxy outcome sub-class	Number	References
analysis transparency	18	1, 3, 6, 11, 21, 26, 40, 41, 45, 46, 48, 49, 50 51, 53, 55, 62, 63
design transparency	2	14, 36
methods transparency general	24	2, 4, 5, 9, 10, 12, 13, 15, 16, 18, 23, 25, 28, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39, 68
research material sharing general	1	17
research process or workflow transparency	6	7, 42, 44, 47, 61, 67

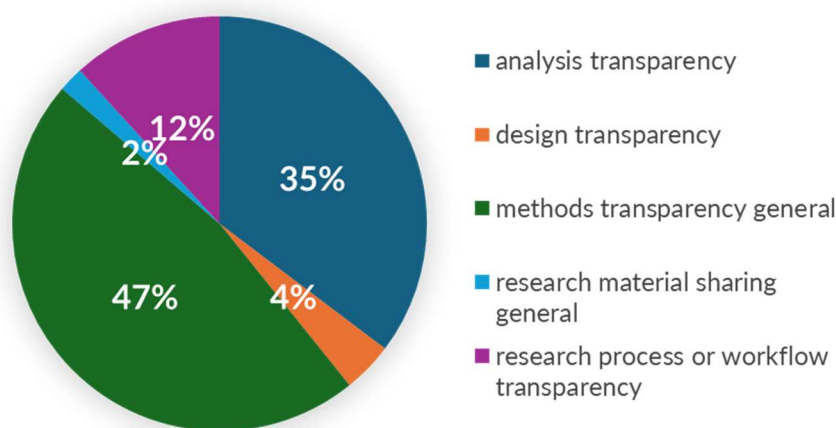


Figure 10: Allocation of interventions per type of proxy outcome sub-class

Figure 10 is a pie chart of Table 9, where quantitative information can be more easily visualized. Almost half of the publications studied interventions whose proxy outcome sub-class is “Methods transparency general”. Another one third of the publications studied interventions whose proxy outcome sub-class is “Analysis transparency”.

3. Delphi survey

3.1 Objective

The aim of this Delphi survey was to create an expert-informed taxonomy of types of interventions for facilitating reproducible research practices at relevant stages of the research cycle within the health sciences, social sciences, and natural sciences. While various interventions exist and have been mapped by the TIER2/OSIRIS and TRUSTparency scoping review, no taxonomy categorizing these interventions currently exists. By generating a consensus taxonomy, we ultimately aimed to provide practical guidance not only for researchers within these fields, but also for institutions, publishers, funders, and other relevant stakeholders seeking to effectively support reproducible research.

3.2 Methodology

To create the expert-informed taxonomy, the Delphi technique was used. The Delphi technique consists of a series of sequential questionnaires or ‘rounds’ where a panel of experts is asked to provide their judgements on a topic (Linstone & Turoff, 1975). After each round, responses are analyzed and fed back to the experts in the next round. The aim of the Delphi method is not to create new knowledge, but to reach consensus among a group of experts (Linstone & Turoff, 1975).

3.2.1 Preparation

In preparation for the first round of the Delphi procedure, we first produced a draft list of stages in the research cycle, along with descriptions of these stages. We also developed a draft list of intervention types and implementing actors, based on a list of interventions gathered by the TIER2/OSIRIS scoping review (Dudda et al., 2025) and supplemented by results from the TRUSTparency scoping review.

Using these lists, Ilse de Jong (AUMc), Joeri Tijndink (AUMc), and Mariska Leeftang (AUMc) constructed a preliminary taxonomy of stages of the research cycle. This initial taxonomy was critically evaluated by other members of the research group (Lex Bouter (AUMc and member of TRUSTparency’s Stakeholder

Advisory Board), Panagiotis Kavouras (UiO) and Mads Sørensen (AU)), who were asked to provide critical feedback and suggest revisions. The revised version of the descriptions was included in the first round of the Delphi study for expert evaluation.

3.2.2 Panelist selection

Panel members were carefully selected for their expertise in the field of reproducibility, meaning that participants were recruited through a purposive sampling technique. TRUSTparenity consortium partners were asked to provide personal contacts of relevant experts. Relevant experts were considered to be: individuals with hands-on experience in researching or implementing reproducibility practices within the fields of health sciences, social sciences, or natural sciences, working at research performing organizations, research funding organizations, publishers, learned societies, and/or embedded in (inter)national reproducibility networks, who are themselves not members of the TRUSTparenity consortium. Since members of our consortium are involved in other reproducibility consortia and networks, this recruitment method ensured we reached a wide list of potential participants. To supplement the list of potential participants, the steering board members of each of the national reproducibility networks were added as potential participants.

3.2.3 Analyses

The Delphi surveys generated both quantitative and qualitative data. Quantitative data consisted of participants' ratings of (1) the descriptions of the (sub)categories within the research cycle, and (2) the relevance of these categories for facilitating reproducible research practices. Ratings were given on a 5-point Likert scale. For agreement with categories and descriptions, a score of 1 indicated strong disagreement and 5 indicated strong agreement. For relevance ratings, 1 reflected low relevance, and 5 reflected high relevance. Consensus was defined as $\geq 67\%$ of participants assigning a score of 4 or 5 to a given category (i.e., agreement or high relevance). Consensus rates were calculated in Excel.

Qualitative data consisted of suggestions and comments experts made concerning the taxonomy of interventions in response to open-ended questions. These responses were coded by Ilse de Jong (AUMc) in Excel by grouping on themes (e.g., suggestion for another example, expression of agreement, suggestion for a change in wording, etc.). The steering committee, consisting of Ilse de Jong (AUMc), Joeri Tjeldink (AUMc), Mariska Leeftink (AUMc), and Lex Bouter (AUMc and member of TRUSTparenity's Stakeholder Advisory Board), discussed the consensus rates and arguments given by panelists to decide on the content of the next round and modifications made to the descriptions and taxonomy structure. The answers given by participants guided the decisions. For example, if consensus was reached, but multiple panelists proposed a similar change to the suggested description of a (sub)category, these changes were considered and proposed in the next round.

3.3 Results

3.3.1 The Delphi panel

In the first round, which was distributed in September of 2025, 228 panelists were invited to participate in the Delphi survey through e-mail. The first round of the survey was also distributed through the LinkedIn pages of the TRUSTparenity and TIER2 projects, and through two invited panelists who offered to share the survey in their network. These non-personal invites led to five extra participants (i.e., 233 in total). Of these five participants, two left their e-mail addresses so they could be contacted for future rounds. The second round of the survey, which was distributed in January of 2026, was only distributed to invited panelists, being the 228 invited to the first round, and the two who left their e-mail address in response to the open survey (i.e., 230 in total).

Seventy-two out of 228 invited panelists participated in the first round of the Delphi (response rate: 31%), and another five participants responded to the openly distributed survey, leading to 77 participants

in the first round. Twenty-three additional invited panelists started but did not complete the survey. Sixteen participants started but did not complete the open survey. As none of these partial completers completed more than half of the survey, we decided not to consider their responses by analyzing the data. In the second round, 73 out of 230 invited panelists participated. Fifty-one of which had also participated in the first round, and the remaining 22 participated only in the second round. In total, 99 out of 230 invited panelists participated in at least one of the rounds (overall response rate: 43%). Table 10 shows the characteristics of the panelists who participated in at least one round. Despite careful geographic distribution of invited panelists, the majority of participating experts work in Europe.

Table 10: Characteristics of panelists who participated in at least one round.

Category	N	Percentage
Gender		
Female	42	42%
Male	56	57%
Non-binary	0	0%
Prefer not to disclose	1	1%
None of the above	0	0%
Geographic region		
Africa	4	4%
Asia	0	0%
Europe	82	83%
Latin America	5	5%
North America	5	5%
Oceania	2	2%
Did not specify	1	1%
Discipline		
Health sciences	46	46%
Natural sciences	14	14%
Social sciences	30	30%
Other	9	9%
Experience with reproducibility		
Not experienced at all	1	1%
Slightly experienced	6	6%
Moderately experienced	35	35%
Very experienced	40	40%
Extremely experienced	17	17%
Role*		
Researcher	80	81%
Research support staff	12	12%
Research manager or administrator	14	14%
Journal editor or publishing professional	16	16%
Member of reproducibility network	52	53%
Member of learned society	8	8%
Policy maker	3	3%
Other	5	5%

*Panelists could select more than one role.

3.3.2 Round 1 Process and Results

In the first round of our Delphi study, we explained to participants that the aim of this Delphi is to develop a taxonomy of interventions that aim to improve reproducible research practices and increase reproducibility. This taxonomy used the stages of the research cycle. Experts were presented with the revised draft version of the taxonomy of stages of the research cycle and with draft descriptions of each stage of the research cycle. Table 11 shows the proposed descriptions for each of the stages presented. Participants were asked to rate the extent to which they agree with each description on a 5-point scale, and to provide comments, suggest improvements, and propose alternative stages, along with justifications for these suggestions. For each stage, participants were asked to rate the relevance for facilitating reproducible research practices on a 5-point scale and were given the option to provide arguments for their rating.

Table 11: Descriptions of research cycle stages as shown in round 1.

Formulate a research question/hypothesis
The early stage in which ideas are formed, the contours of the study are described and resources are secured. Includes: - Specify the research question/hypotheses: Develop theoretically or empirically grounded research questions/hypotheses. - Sketch the contours of the study plan. - Obtain resources: Preparing applications to secure financial, human, and technical resources.
Design of the study
The planning stage in which methodological decisions are made, operationalized, and documented. Includes: - Develop study protocol: Creating a detailed plan for how the study will be conducted, including sampling, measurement instruments, data-analysis plan, and data management plan. - Operationalize the study protocol: Obtaining or creating the materials, instruments, tools, or software needed to execute the study. - Uploading the study protocol in a time-stamped repository before the start of data collection.
Data collection
The execution stage of the research, involving data collection, storage, and documentation. Includes: - Collect data: Executing the data-gathering process (e.g., running experiments, conducting interviews, administering surveys). - Store data. - Document procedures: Keeping detailed records of how the study was conducted, including any deviations from the study protocol and decisions made during execution.
Data analysis
The stage where collected data is prepared, processed, and analyzed to answer the research question. Includes: - Prepare and clean data: Organize, transform, and clean raw data to prepare it for analysis. - Run analyses: Conduct statistical or qualitative analyses. - Validate results: Assess the robustness and reliability of findings.
Report the findings
The stage where findings are interpreted and shared with the (scholarly) community. Includes: - Interpret findings: Contextualize the results in relation to the research question, hypothesis, theory, and previous literature. - Write the report: Report the findings of the study. - Publish/disseminate findings: Submit the manuscript to a preprint server, journal or conference.

- Share data/code/materials: Make datasets, analysis scripts, and research materials available for accountability or reuse purposes.

Post-publication

The post-publication stage in which findings are assessed, scrutinized, and extended. Includes:

- Peer-review and commentary: Participating in or responding to peer review processes and critically assess the study's design, analysis, and interpretation.
- Corrections: Addressing errors, omissions, or limitations discovered after publication by issuing corrections, retractions, or clarifications.
- Replication: Conducting independent attempts to reproduce the findings of the study using the same or closely related methods and materials.

For all presented items, consensus was reached based on our predefined rate of consensus of 67% of panelists strongly agreeing or agreeing with a question or rating the relevance as very or extremely relevant. Rates of consensus ranged from 71% to 97%. Table 12 shows all ratings of consensus and number of comments.

Table 12: Rates of consensus and number of comments for all stages of the research cycle presented in round 1.

Stage	Rate of consensus		Number of comments		
	Description	Relevance	Changes to description	Addition to description	Arguments for relevance
Formulate a RQ	71%	71%	32	33	34
Design study	82%	95%	29	27	24
Collect data	86%	94%	24	24	20
Analyze data	88%	97%	25	21	17
Write report	82%	92%	22	20	24
Post-publication	79%	71%	28	18	19

On average, we received 24 comments for each open-ended question that we asked. All comments were analyzed and grouped by type of answer and overlapping content between answers by Ilse de Jong (AUmc) and checked by Joeri Tjink (AUmc) and Mariska Leeflang (AUmc). Summaries of the responses given to each question and selected example comments were included in a feedback report that was presented to all participants for round 2. The Delphi round 1 survey is included in Annex 5.8 and the full feedback report for round 1 is included in Annex 5.9.

3.3.3 Round 2 Process and Results

Based on the consensus rates and comments given in the first round of the Delphi, the taxonomy for the second round was prepared (see Annex 5.11). Even though consensus was reached for all presented categories in the first round, frequently occurring comments were used to inform changes made to the taxonomy in the second round. This led to three main changes and adjustments to the descriptions of (sub)categories to make them more general and applicable to a variety of research contexts. The three main changes included:

- 1) the merging of "Formulate a research question/hypothesis" and "Design of the study" into one category: "Preparation"
- 2) the addition of an "Infrastructure" category
- 3) the re-naming of the "Write the report" stage to "Disseminate", and the addition of "Post-publication" and "Peer-review" as subcategories to this stage.

In round 2, participants were asked to rate their agreement to each of the proposed main changes on a scale of 1 (*Completely disagree*) to 5 (*Completely agree*), and to provide their arguments for their rating. Additionally, participants were presented with the updated descriptions for the subcategories of each of the stages of the research cycle. All subcategories were updated with examples of potential interventions that could be included in each category. At the end of each overarching category, panelists were asked if there are any interventions that they believe cannot be placed into any of the presented subcategories and were given the option to suggest additional subcategories. Table 13 shows the descriptions of the (sub)categories as presented in round 2. The full round 2 survey can be found in Annex 5.10.

Table 13: Descriptions of (sub)categories as presented in round 2.

Preparation
Formulate the research question/hypothesis The process of defining what the study aims to investigate. Examples of interventions that could fall into this category are: conducting literature reviews; using reporting guidelines or checklists for formulating research questions/hypotheses; etc.
Plan the study Creating a plan for how the study will be conducted. Examples of interventions that could fall into this category are: developing a study protocol; writing a data management plan; writing a statistical analysis plan; preregistration; pilot testing; etc.
Obtaining resources Securing the resources necessary to execute the study. Examples of interventions that could fall into this category are: assembling or training the research team; obtaining or creating materials; requesting funding; etc.
Data collection
Data collection Executing the data-gathering process. Examples of interventions that could fall into this category are: standardization of data collection procedures; training in data collection methods; etc.
Data storage Saving and organizing the collected data. Examples of interventions that could fall into this category are: using secure data repositories; use of data management procedures; etc.
Document procedures Recording details of how the data are collected. Examples of interventions that could fall into this category are: maintaining lab notebooks; documenting changes to the original protocol; etc.
Data analysis
Data preparation Organizing and cleaning the data to prepare for analysis. Examples of interventions that could fall into this category are: converting data into interoperable formats; using controlled vocabularies or coding schemes to standardize variables; etc.
Conduct analyses Conducting quantitative and/or qualitative analysis. Examples of interventions that could fall into this category are: use of open source analysis software; applying standardized analysis methods; etc.
Data validation Assessing the accuracy and reliability of findings. Examples of interventions that could fall into this category are: conducting robustness checks; independent re-analysis of results; data integrity checks; using an effect-size filter; etc.
Document procedures

Recording details of how the data are analyzed. Examples of interventions that could fall into this category are: maintaining an analysis log or notebook; using version controlling for data analysis; etc.

Dissemination

Write the report

Report the findings of the study. Examples of interventions that could fall into this category are: complete reporting; using a reporting guideline; etc.

Share data

Making research data accessible to others. Examples of interventions that could fall into this category are: sharing data in repositories; applying the FAIR-principles; etc.

Peer-review

The evaluation of research by others. Examples of interventions that could fall into this category are: use of reproducibility checklist during peer-review; guideline reminders for peer-reviewers; etc.

Post-publication

Contributing to the ongoing use, evaluation, and discussion of research after publication. Examples of interventions that could fall into this category are: writing letters to editors or authors to highlight issues; engaging in replication studies; responding to requests to re-use data; etc.

Infrastructure

Research performing organizations (RPOs)

Organizations that perform research and provide resources, policies, and support for their researchers. Examples of interventions that could fall into this category are: offering training and support in open science and/or reproducible research; implementing policies on data management; recognizing and rewarding reproducible research practices in hiring and promotion; etc.

Research funding organizations (RFOs)

Organizations that fund research and set requirements or incentives for reproducibility. Examples of interventions that could fall into this category are: requiring data management plans and/or open data sharing as part of grant applications; funding infrastructure for open science/reproducibility; mandating preregistration for funded projects; etc.

Publishers

Organizations that disseminate research and set standards for transparency and reproducibility in publication. Examples of interventions that could fall into this category are: requiring data and/or code availability statements in publications; offering registered reports options; implementing reproducibility checklists; recognizing and rewarding reproducible research practices through badges; etc.

Consensus was reached for all but the first main change - i.e. the merging of "Formulate a research question/hypothesis" and "Design of the study" into one category: "Preparation" - and for all presented subcategories. Consensus rates and number of comments are shown in Table 14.

Table 14: Rates of consensus and number of comments for proposed main changes in round 2.

Proposed change	Rate of consensus	Number of comments
To what extent do you agree with the proposal to merge "Formulate a research question/hypothesis" and "Design of the study" into one category: "Preparation"?	56%	38
To what extent do you agree with the proposal to add an "Infrastructure" category?	81%	33
To what extent do you agree with the proposal to re-name the "Write the report" stage to "Disseminate", and to add "Post-publication" and "Peer-review" as subcategories to this stage?	74%	26

Table 15: Rates of consensus and number of comments for all stages of the research cycle presented in round 2.

Category	Subcategory	Rate of consensus	Number of comments
Preparation	• Formulate the RQ/hypothesis	81%	27
	• Plan the study	89%	21
	• Obtaining resources	74%	23
Data collection	• Data collection	90%	15
	• Data storage	90%	16
	• Document procedures	85%	21
Data analysis	• Data preparation	89%	14
	• Conduct analyses	90%	14
	• Data validation	81%	16
	• Document procedures	86%	19
Dissemination	• Write the report	93%	18
	• Share data	82%	23
	• Peer-review	77%	24
	• Post-publication	79%	18
Infrastructure	• RPOs	84%	15
	• RFOs	85%	16
	• Publishers	77%	21

In response to the first main change (i.e., the proposal to merge “Formulate the research question/hypothesis” and “Design of the study” into “Preparation”), for which consensus was not reached, 14 panelists commented that the two stages are fundamentally different, with “Formulate the research question” relating to the “what” of research, and “Design of the study” relating to the “how” of research. Five panelists made a similar comment but argued that they would agree with merging these two categories, considering that the two are presented as distinct subcategories within the “Preparation” stage, meaning that the different nature of the two stages is still recognized. Eleven participants expressed their agreement with the proposal to merge the categories.

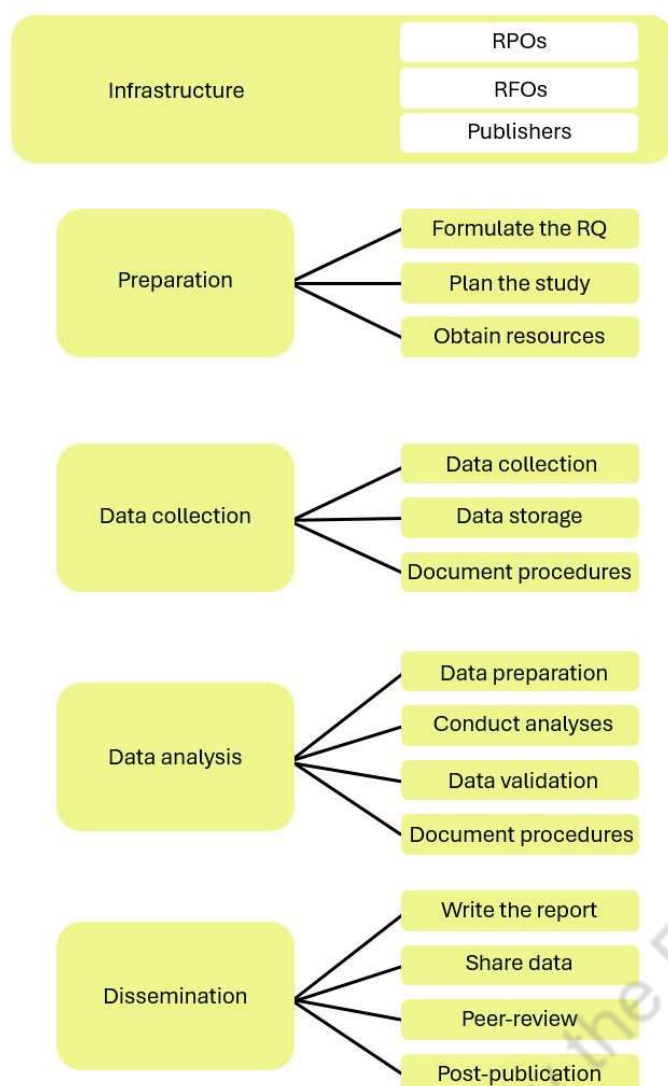


Figure 11: Expert-informed taxonomy of reproducibility interventions after round 2.

Considering that the originally distinct categories are still presented as distinct subcategories in the updated taxonomy, and considering that consensus was reached for all of the proposed subcategories within the “Preparation” stage, we decided to keep the taxonomy as proposed in round 2 unchanged for the placement of interventions to the different categories within the interventions, despite not reaching consensus on this proposed main change. The taxonomy after round 2 of the Delphi survey is presented in **Figure 11**.

Categorization of interventions

All interventions identified through the TIER2/OSIRIS and the TRUSTparenacy scoping reviews were categorized into the subcategories created through the Delphi consensus survey. This process was executed by Pooja Aryal (UiO) and Ilse de Jong (AUMc). Interventions for which categorization was not agreed upon were categorized by Panagiotis Kavouras (UiO) and Joeri Tjink (AUMc). Table 16 shows all interventions categorized in the expert-informed taxonomy. The numbering corresponds to the one presented in Annex 5.11.

The majority of interventions exists in the infrastructure category (72), although these mostly consist of interventions implemented by journals (66), specifically relating to policy implementation (51). After infrastructure, the

data analysis stage of the research cycle is most populated by tested interventions (55). Interestingly, the data analysis stage of the research cycle also scored highest in terms of consensus on the relevance of this stage for facilitating reproducible research practice (97% consensus). The least number of interventions exist in the data collection stage, with only 14 interventions categorized into one of its subcategories. Twenty-five interventions were categorized into the preparation stage, and 44 interventions were categorized into the dissemination stage. The “formulate the research question/hypothesis” subcategory was the only subcategory not to include any interventions.

Table 16: Categorization of reproducibility interventions.

Main category	Subcategory	Intervention	Individual studies
Infrastructure	Research performing organizations (RPOs)	open science role model in supervision	24, 114
		providing support in RDM	146
		funder policy on trial registration	154*

	Research funding organizations (RFOs)	funder policy on data sharing	143*, 144
	Publishers	journal policy on using reporting checklist	58, 69, 70, 82*, 85, 86, 101, 115, 172, 173
		journal policy on data/code/materials sharing	8, 71, 73, 74, 75, 77, 78*, 89, 90, 92, 103, 108, 112, 123, 124, 127, 132, 133, 134, 135, 137, 138, 140, 143*, 145, 147, 151, 153, 156, 163, 168
		journal policy on promoting computational reproducibility	12
		journal editor checks completeness of reporting	79*
		journal accepts regardless of outcome	80
		badge for open data/code	88, 113, 119, 148, 149, 150
		journal submission guidelines	96, 98, 107, 109, 121, 122
		badge for open science	97, 99, 139
		journal policy on trial registration	110, 111, 129, 154*, 157, 158
		results-free peer-review	94
Preparation	Formulate the RQ		
	Plan the study	registered reports	141, 160
		statistical tool to estimate power	23
		development and implementation of SOPs	5, 16*, 21, 30, 38*, 40*, 41, 45*, 67*, 72*, 155
		preregistration/publication of protocol	31*, 100, 164, 165, 166, 169
	Obtain resources	training on implementing SOPs	38
		training on computing skills	91*
		training on doing a systematic review	130
		use of toolbox to create standardized experiments	54*
	Development of open automated quantification framework		14
Data collection	Data collection	implementation of SOPs	16*, 22, 29, 38*, 44, 67*, 72*, 93
		collection of data in standardized formats	52, 54*
	Data storage	structured framework to organize projects	62*
	Document procedures	use of a structured reporting document	65*, 68*
		standardized R package for data retrieval	11
		real-time reporting	31*
Data analysis	Data preparation	R function to convert datasets for efficient analysis	1, 36*
		implementation of standardized data processing protocol	5*, 9, 16*, 26*, 35*, 40*, 42*, 47, 64, 65*
		method that normalises datasets and computes local consistency	35*, 48, 49, 57, 59, 63, 118
		training on computing skills	91*

Dissemination	Conduct analyses	bias adjustment method for dataset	3
		implementation of standardized data processing protocol	5*, 6*, 7*, 26*, 35*, 36*, 37*, 39*, 40*, 42*, 46*, 56*, 61
	Data validation	statistical post-test effect size filter	104
		standardized testing framework enabling reliable comparison of results across platforms	10, 17, 36*
		quality assessment for genome data	13, 15, 55
		heterogenization test in multiple labs	28, 60
		implementation of standardized framework to evaluate quality	6*, 7*, 39*, 40*, 56*
		methodological and statistical intervention for causal inference that introduces simulation-based evaluation to compare estimators	45
		tool to select the most representative model from multiple runs	51
		multiverse analysis to assess robustness of analytic decisions	50
	Document procedures	use of a structured reporting document	40*, 42*, 68*
	Write the report	web-based writing tool	76, 116, 162
		using a reporting guideline/checklist	81, 82*, 83, 93, 95, 117, 125, 131, 136, 142, 152, 170, 171
	Share data	sharing data	27, 31*, 78*
		data curation tool in accordance with FAIR	25, 33, 53, 126*, 159
		tool for evaluation of FAIRness	2, 19
		using privacy-preserving approach for data-sharing	4
		standardized open-data structure	32, 62*
		R-based intervention converting scripts into standardized workflows	37*
		open-source code repository	61
	Peer-review	journal editor/peer-reviewer checks completeness of reporting	79*, 87, 161
		results-free peer-review	94
		statistical tool to estimate reproducibility	84
	Post-publication	sending correction letters to journal editors	102
		post-publication p-value calibration	105, 106
		writing e-mails to request data	34, 120*
		responding to requests to share data	120*
		implementation of platform to assess replicability of scientific claims	18
		automatic code cleaning used post-publication	167

*interventions included in multiple categories.

3.4 Next steps

All interventions identified through the TIER2/OSIRIS and TRUSTparency scoping reviews and categorized into this expert-informed taxonomy, will all be evaluated based on parameters relating to strength of the evidence and implementability to further shape the TRUSTparency toolbox of interventions. This line of work will be performed in the context of Task 3.2.

4. References

Dudda L, Kormann E, Kozula M, DeVito NJ, Klebel T, Dewi APM, Spijker R, Stegeman I, Van den Eynden V, Ross-Hellauer T, Leeflang MMG. Open science interventions to improve reproducibility and replicability of research: a scoping review. *R Soc Open Sci.* 2025 Apr 9;12(4):242057. doi: 10.1098/rsos.242057. PMID: 40206851; PMCID: PMC11979971.

Linstone H, Turoff M. The Delphi Method: Techniques and Applications. *J Mark Res.* 1975 18(3) DOI: 10.2307/3150755.

Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med.* 2018;169: 467–473. doi:10.7326/M18-0850.

Under review by the European Commission

5. Annexes

5.1 Search strategies

EMBASE through Ovid: Embase Classic+Embase <2023 August 14 - today>

#	Searches
1	((data or code or workflow or practices or materials or notebook) adj2 (open or share or shared or sharing or preservation or stewardship)) or "open science" or ((computational or data or open or research or conclusion* or inferential or analytic or conceptual or direct or exact or statistical) adj3 (reproducib* or replicability or replicable)) or (research adj5 (transparen* or credib*)))ti,ab. or (reporting adj3 guideline*).ti.
2	Randomized Controlled Trial/ or Controlled Clinical Trial/ or Quasi Experimental Study/ or Pretest Posttest Control Group Design/ or Time Series Analysis/ or Experimental Design/ or Multicenter Study/ or random*.ti,ab. or groups.ab. or (trial or multicentre or multicenter or multi centre or multi center).ti. or (intervention? or effect? or impact? or controlled or control group? or (before adj12 after) or (pre adj5 post) or ((pretest or pre test) and (posttest or post test)) or quasiexperiment* or quasi experiment* or pseudo experiment* or pseudoexperiment* or evaluat* or time series or time point? or repeated measur* or ((experimental or empirical or qualitative) adj5 (study or studies))).ti,ab.
3	exp clinical study/ or (Cohort adj (study or studies)).mp. or (Case control adj (study or studies)).tw. or (follow up adj (study or studies)).tw. or (observational adj (study or studies)).tw. or (epidemiologic\$ adj (study or studies)).tw. or (cross sectional adj (study or studies)).tw. or (comparative adj stud*).mp.
4	2 or 3
5	1 and 4
6	limit 5 to embase

Medline: Ovid MEDLINE(R) ALL <2023 August 14 - today>

#	Searches
1	((data or code or workflow or practices or materials or notebook) adj2 (open or share or shared or sharing or preservation or stewardship)) or "open science" or ((computational or data or open or research or conclusion* or inferential or analytic or conceptual or direct or exact or statistical) adj3 (reproducib* or replicability or replicable)) or (research adj5 (transparen* or credib*)))ti,ab. or (reporting adj3 guideline*).ti.

2	(Clinical study/ or Case control study/ or Family study/ or Longitudinal study/ or Retrospective study/ or Prospective study/ or Cohort analysis/ or Comparative Study/ or (Cohort adj (study or studies)).mp. or (Case control adj (study or studies)).tw. or (follow up adj (study or studies)).tw. or (observational adj (study or studies)).tw. or (epidemiologic\$ adj (study or studies)).tw. or (cross sectional adj (study or studies)).tw. or (comparative adj stud*).mp. or (("randomized controlled trial" or "controlled clinical trial" or "multicenter study" or "pragmatic clinical trial").pt. or non-randomized controlled trials as topic/ or interrupted time series analysis/ or controlled before-after studies/ or random*.ti,ab. or groups.ab. or (trial or multicenter or "multi center" or multicentre or "multi centre").ti. or (intervention? or effect? or impact? or controlled or control group? or (before adj12 after) or (pre adj5 post) or ((pretest or "pre test") and (posttest or "post test")) or quasiexperiment* or quasi experiment* or pseudo-experiment* or pseudoexperiment* or evaluat* or "time series" or time-point? or "repeated measur*" or ((experimental or empirical or qualitative) adj5 (study or studies))).ti,ab.)) not ((news or comment or editorial).pt. or comment on.cm.)
3	1 and 2

ERIC <2023 August 1 - today>

#	Searches
1	((data or code or workflow or practices or materials or notebook) adj2 (open or share or shared or sharing or preservation or stewardship)) or "open science" or ((computational or data or open or research or conclusion* or inferential or analytic or conceptual or direct or exact or statistical) adj3 (reproducib* or replicability or replicable)) or (research adj5 (transparen* or credib*))).ti,ab. or (reporting adj3 guideline*).ti.
2	(Clinical study/ or Case control study/ or Family study/ or Longitudinal study/ or Retrospective study/ or Prospective study/ or Cohort analysis/ or Comparative Study/ or (Cohort adj (study or studies)).mp. or (Case control adj (study or studies)).tw. or (follow up adj (study or studies)).tw. or (observational adj (study or studies)).tw. or (epidemiologic\$ adj (study or studies)).tw. or (cross sectional adj (study or studies)).tw. or (comparative adj stud*).mp. or (("randomized controlled trial" or "controlled clinical trial" or "multicenter study" or "pragmatic clinical trial").pt. or non-randomized controlled trials as topic/ or interrupted time series analysis/ or controlled before-after studies/ or random*.ti,ab. or groups.ab. or (trial or multicenter or "multi center" or multicentre or "multi centre").ti. or (intervention? or effect? or impact? or controlled or control group? or (before adj12 after) or (pre adj5 post) or ((pretest or "pre test") and (posttest or "post test")) or quasiexperiment* or quasi experiment* or pseudo-experiment* or pseudoexperiment* or evaluat* or "time series" or time-point? or "repeated measur*" or ((experimental or empirical or qualitative) adj5 (study or studies))).ti,ab.)) not (news or comment or editorial).pt.
3	1 and 2

APA PsycInfo <August Week 1 2023 - today>

#	Searches
---	----------

1	((data or code or workflow or practices or materials or notebook) adj2 (open or share or shared or sharing or preservation or stewardship)) or "open science" or ((computational or data or open or research or conclusion* or inferential or analytic or conceptual or direct or exact or statistical) adj3 (reproducib* or replicability or replicable)) or (research adj5 (transparen* or credib*))).ti,ab. or (reporting adj3 guideline*).ti.
2	(Clinical study/ or Case control study/ or Family study/ or Longitudinal study/ or Retrospective study/ or Prospective study/ or Cohort analysis/ or Comparative Study/ or (Cohort adj (study or studies)).mp. or (Case control adj (study or studies)).tw. or (follow up adj (study or studies)).tw. or (observational adj (study or studies)).tw. or (epidemiologic\$ adj (study or studies)).tw. or (cross sectional adj (study or studies)).tw. or (comparative adj stud*).mp. or (("randomized controlled trial" or "controlled clinical trial" or "multicenter study" or "pragmatic clinical trial").pt. or non-randomized controlled trials as topic/ or interrupted time series analysis/ or controlled before-after studies/ or random*.ti,ab. or groups.ab. or (trial or multicenter or "multi center" or multicentre or "multi centre").ti. or (intervention? or effect? or impact? or controlled or control group? or (before adj12 after) or (pre adj5 post) or ((pretest or "pre test") and (posttest or "post test")) or quasiexperiment* or quasi experiment* or pseudo-experiment* or pseudoexperiment* or evaluat* or "time series" or time-point? or "repeated measur*" or ((experimental or empirical or qualitative) adj5 (study or studies))).ti,ab.)) not (news or comment or editorial).pt.
3	1 and 2

Scopus (August 2023-today)

1	(TITLE-ABS (((data OR code OR workflow OR practices OR materials OR notebook) W/1 (share OR shared OR sharing OR preservation OR stewardship)) OR "open science" OR ((computational OR data OR open OR research OR conclusion* OR inferential OR analytic OR conceptual OR direct OR exact OR statistical) W/2 (reproducib* OR replicability OR replicable)) OR (research W/2 (transparen* OR credib*))) OR TITLE (reporting W/2 guideline*)) AND (TITLE-ABS (cohort W/1 (study OR studies)) OR TITLE-ABS ("Case control" W/1 (study OR studies)) OR TITLE-ABS (follow-up W/1 (study OR studies)) OR TITLE-ABS (observational W/1 (study OR studies)) OR TITLE-ABS ("cross sectional" W/1 (study OR studies)) OR TITLE-ABS (comparative W/1 stud*) OR TITLE-ABS (random*) OR TITLE (multicenter OR "multi center" OR multicentre OR "multi centre") OR TITLE-ABS (intervention* OR effectiveness OR (controlled W/3 trial*) OR "control group*" OR (before W/11 after) OR ((pretest OR "pre test") AND (posttest OR "post test")) OR quasiexperiment* OR "quasi experiment*" OR "pseudo experiment*" OR pseudoexperiment* OR evaluation OR ((experimental OR empirical OR qualitative) W/4 (study OR studies)))) AND (EXCLUDE (DOCTYPE , "cp"))
---	---

Web of Science (August 2023-today)

1	((TI= (((data OR code OR workflow OR practices OR materials OR notebook) NEAR/1 (share OR shared OR sharing OR preservation OR stewardship)) OR "open science" OR ((computational OR data OR open OR research OR conclusion* OR inferential OR analytic OR conceptual OR direct OR exact OR statistical) NEAR/2 reproducib*) OR (research NEAR/2 (transparen* OR credib*)) OR (reporting NEAR/2 guideline*))) AND (TI= (cohort NEAR/1 (study OR studies)) OR TI= ("Case control" NEAR/1 (study OR studies)) OR TI= (follow-up NEAR/1 (study OR studies)) OR TI= (observational NEAR/1 (study OR studies)) OR TI= ("cross sectional" NEAR/1 (study OR studies)) OR TI= (comparative NEAR/1 stud*) OR TI= (random*) OR TI= (multicenter OR "multi center" OR multicentre OR "multi centre") OR TI= (intervention* OR effectiveness OR (controlled NEAR/3 trial*) OR "control group*" OR (before NEAR/12 after) OR ((pretest OR "pre test") AND (posttest OR "post test"))) OR quasiexperiment* OR "quasi experiment*" OR "pseudo experiment*" OR pseudoexperiment* OR evaluation OR ((experimental or empirical or qualitative) NEAR/4 (study or studies)))) OR ((AB= (((data OR code OR workflow OR practices OR materials OR notebook) NEAR/1 (share OR shared OR sharing OR preservation OR stewardship)) OR "open science" OR ((computational OR data OR open OR research OR conclusion* OR inferential OR analytic OR conceptual OR direct OR exact OR statistical) NEAR/2 reproducib*) OR (research NEAR/2 (transparen* OR credib*))))) AND (AB= (cohort NEAR/1 (study OR studies)) OR AB= ("Case control" NEAR/1 (study OR studies)) OR AB= (follow-up NEAR/1 (study OR studies)) OR AB= (observational NEAR/1 (study OR studies)) OR AB= ("cross sectional" NEAR/1 (study OR studies)) OR AB= (comparative NEAR/1 stud*) OR AB= (random*) OR AB= (multicenter OR "multi center" OR multicentre OR "multi centre") OR AB= (intervention* OR effectiveness OR (controlled NEAR/3 trial*) OR "control group*" OR (before NEAR/12 after) OR ((pretest OR "pre test") AND (posttest OR "post test"))) OR quasiexperiment* OR "quasi experiment*" OR "pseudo experiment*" OR pseudoexperiment* OR evaluation OR ((experimental or empirical or qualitative) NEAR/4 (study or studies))))))
---	---

5.2 Inclusion and exclusion criteria

Original criterion	New criterion
Study characteristics	
Include - Studies evaluating the effectiveness of interventions aimed at enhancing the reproducibility or replicability of scientific methods and findings in a comparative design. These could be either between-subject or within-subject comparisons. - Non-comparative studies were only included if the introduction of an intervention was explicitly stated (e.g. post-intervention study reporting the level of reproducibility after the implementation of a guideline).	No change
Exclude - Studies if they did not test an intervention; if the full text was unavailable; and if the article did not report primary resources, such as Stage 1 registered reports or position papers. - Studies that investigated agreement in test results between observers or laboratories, which is also sometimes referred to as reproducibility of test results. - Studies that only report the prevalence of an intervention (e.g., measuring only the amount of preregistration in a field).	No change
Population	
Include - Researchers of any career level - Holders of other academic posts such as editors or administrative staff - An entire academic institution such as a university, journal, publisher or funder. - The included studies could cover any scientific field or discipline.	No change
Exclude N/A	No change
Intervention/Exposure	
Include Any intervention that aimed to improve the reproducibility and replicability of science, basing the categories of interventions on broad definitions taken from multiple disciplines.	Include Any study that includes both elements: interventions and assessment of the outcome, in terms of reproducibility and/or replicability. For example, a study that assesses interventions to improve reproducibility (e.g. whether pre-registered studies yield similar results in replication) or to improve replicability (e.g. an

	intervention to improve adherence to reporting guidelines or to improve pre-registration can count as an intervention).
Exclude N/A	Exclude Any study that includes interventions without assessment of the outcome, in terms of reproducibility and/or replicability. For example, we do not include studies that assess the number of studies that follow a particular reporting guideline, or studies that replicate studies to see how reproducible they are.
Outcome	
Include - The reproducibility or replicability of research. This means that studies attempt to replicate or reproduce another study, including aspects such as data collection, analyses, and/or conclusions. - Proxy outcomes that would be necessary, but not necessarily sufficient, for improved reproducibility (e.g. preregistration, data sharing, and completeness of reporting).	Include - This is going to become an exclusion criterion (see line below). - This is going to be removed.
Exclude N/A	Exclude Studies that attempt to reproduce or replicate another study, including aspects such as data collection, analyses, and/or conclusions, without testing an intervention.
Other	
Include Any article type containing primary data, including peer reviewed original articles, early access papers and preprints.	No change
Exclude - Position papers, study protocols, editorials and any other articles lacking primary data. - Reviews.	No change

5.3 Articles for data extraction – bibliographic records in alphabetical order

(surname of first author)

1. Abdelaziz AI, Hanson KA, Gaber CE, Lee TA. Optimizing large real-world data analysis with parquet files in R: A step-by-step tutorial. *Pharmacoepidemiol Drug Saf.* 2024 Mar;33(3):e5728. doi: 10.1002/pds.5728. Epub 2023 Nov 20. PMID: 37984998.
2. Aguilar Gómez F, Bernal I. FAIR EVA: Bringing institutional multidisciplinary repositories into the FAIR picture. *Sci Data.* 2023 Nov 4;10(1):764. doi: 10.1038/s41597-023-02652-8. PMID: 37925573; PMCID: PMC10625599.
3. Ahern TP, Collin LJ, MacLehose RF, Littenberg B, Haines L, Bonnett M, Asmussen FB, Chen J, Lash TL. Adjusting Adjustments: Using External Data to Estimate the Impact of Different Confounder Sets on Published Associations. *Epidemiology.* 2025 May 1;36(3):381-390. doi: 10.1097/EDE.0000000000001821. Epub 2024 Nov 22. PMID: 39575936; PMCID: PMC11991891.
4. Anido-Alonso, A., & Alvarez-Estevez, D. (2023). Decentralized Data-Privacy Preserving Deep-Learning Approaches for Enhancing Inter-Database Generalization in Automatic Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*, 27, 5610–5621. <https://doi.org/10.1109/jbhi.2023.3310869>.
5. Ashkarran AA, Gharibi H, Modaresi SM, Saei AA, Mahmoudi M. Standardizing Protein Corona Characterization in Nanomedicine: A Multicenter Study to Enhance Reproducibility and Data Homogeneity. *Nano Lett.* 2024 Aug 14;24(32):9874-9881. doi: 10.1021/acs.nanolett.4c02076. Epub 2024 Aug 3. PMID: 39096192; PMCID: PMC11328176.
6. Bahl, Aileen & Ibrahim, Celine & Plate, Kristina & Haase, Andrea & Dengjel, Joern & Nymark, Penny & Dumit, Verónica. (2023). PROTEOMAS: a workflow enabling harmonized proteomic meta-analysis and proteomic signature mapping. *Journal of Cheminformatics*. 15. 10.1186/s13321-023-00710-2.
7. Bontempi D, Nuernberg L, Pai S, Krishnaswamy D, Thiriveedhi V, Hosny A, Mak RH, Farahani K, Kikinis R, Fedorov A, Aerts HJWL. End-to-end reproducible AI pipelines in radiology using the cloud. *Nat Commun.* 2024 Aug 13;15(1):6931. doi: 10.1038/s41467-024-51202-2. PMID: 39138215; PMCID: PMC11322541.
8. Bramley, Paul. (2024). Ask, and it shall be given you – individual patient data and code availability for randomised controlled trials submitted for publication. *Anaesthesia*. 80. 10.1111/anae.16503.
9. Chen YW, Davenport O, Yu N, Melendez R, Nugteren J, Nijem I, Yan W, Hendricks R, Song Y. Addressing Clinical Challenges in Aberrant Pharmacokinetics of Biologic Therapeutic Drugs: Investigating Sample Processing Procedure in the Immunoassays. *AAPS J.* 2025 Feb 18;27(2):46. doi: 10.1208/s12248-025-01036-7. PMID: 39966314.
10. Denisova, Elena & Tiribilli, Eleonora & Luschi, Alessio & Francia, Piergiorgio & Bocchi, Leonardo & Manetti, Leonardo & Iadanza, Ernesto. (2024). Enabling reliable usability assessment and comparative analysis of medical software: a comprehensive framework for multimodal biomedical imaging platforms. *Health and Technology*. 14. 1-12. 10.1007/s12553-024-00859-2.
11. de Vries P. ECOTOXr: An R package for reproducible and transparent retrieval of data from EPA's ECOTOX database. *Chemosphere.* 2024 Sep;364:143078. doi: 10.1016/j.chemosphere.2024.143078. Epub 2024 Aug 23. PMID: 39181462.
12. Ellis DA, Towse J, Brown O, Cork A, Davidson BI, Devereux S, Hinds J, Ivory M, Nightingale S, Parry DA, Piwek L, Shaw H, Towse AS. Assessing computational reproducibility in Behavior Research Methods. *Behav*

- Res Methods. 2024 Dec;56(8):8745-8760. doi: 10.3758/s13428-024-02501-5. Epub 2024 Sep 25. PMID: 39322919; PMCID: PMC11525395.
13. Elmanzalawi M, Fujisawa T, Mori H, Nakamura Y, Tanizawa Y. DFAST_QC: quality assessment and taxonomic identification tool for prokaryotic Genomes. BMC Bioinformatics. 2025 Jan 7;26(1):3. doi: 10.1186/s12859-024-06030-y. PMID: 39773409; PMCID: PMC11705978.
 14. Fernholz MHP, Guggiana Nilo DA, Bonhoeffer T, Kist AM. DeepD3, an open framework for automated quantification of dendritic spines. PLoS Comput Biol. 2024 Feb 29;20(2):e1011774. doi: 10.1371/journal.pcbi.1011774. PMID: 38422112; PMCID: PMC10903918.
 15. Ferreira MJ, Silva J, Pinto SC, Coimbra S. I Choose You: Selecting Accurate Reference Genes for qPCR Expression Analysis in Reproductive Tissues in *Arabidopsis thaliana*. Biomolecules. 2023 Mar 2;13(3):463. doi: 10.3390/biom13030463. PMID: 36979397; PMCID: PMC10046263.
 16. Fiedler S, Schrader H, Theobalt N, Hofmann I, Geiger T, Arndt D, Wanke R, Schwaiger J, Blutke A. Standardized tissue sampling guidelines for histopathological and molecular analyses of rainbow trout (*Oncorhynchus mykiss*) in ecotoxicological studies. PLoS One. 2023 Jul 13;18(7):e0288542. doi: 10.1371/journal.pone.0288542. PMID: 37440561; PMCID: PMC10343068.
 17. Ernestas Filatovas, Linas Stripinis, Francisco Orts, Remigijus Paulavičius, *Advancing Research Reproducibility in Machine Learning through Blockchain Technology*, Informatica 35(2024), no. 2, 227-253, DOI 10.15388/24-INFOR553.
 18. Fraser H, Bush M, Wintle BC, Mody F, Smith ET, Hanea AM, Gould E, Hemming V, Hamilton DG, Rumpff L, Wilkinson DP, Pearson R, Singleton Thorn F, Ashton R, Willcox A, Gray CT, Head A, Ross M, Groenewegen R, Marcoci A, Vercammen A, Parker TH, Hoekstra R, Nakagawa S, Mandel DR, van Ravenzwaaij D, McBride M, Sinnott RO, Vesk P, Burgman M, Fidler F. Predicting reliability through structured expert elicitation with the replicATS (Collaborative Assessments for Trustworthy Science) process. PLoS One. 2023 Jan 26;18(1):e0274429. doi: 10.1371/journal.pone.0274429. PMID: 36701303; PMCID: PMC9879480.
 19. Gaignard A, Rosnet T, De Lamotte F, Lefort V, Devignes MD. FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. J Biomed Semantics. 2023 Jul 1;14(1):7. doi: 10.1186/s13326-023-00289-5. PMID: 37393296; PMCID: PMC10315041.
 20. Gierend K, Waltemath D, Ganslandt T, Siegel F. Traceable Research Data Sharing in a German Medical Data Integration Center With FAIR (Findability, Accessibility, Interoperability, and Reusability)-Geared Provenance Implementation: Proof-of-Concept Study. JMIR Form Res. 2023 Dec 7;7:e50027. doi: 10.2196/50027. PMID: 38060305; PMCID: PMC10739241.
 21. Gillies MB, Bharat C, Camacho X, Daniels B, Hall K, Kelly TL, Kelty E, Lin J, Litchfield M, Lopez D, Nogrehchi F, Raubenheimer J, Varney B, Pratt N. Medicine utilization studies in Australian individual-level dispensing data: A blinded, multi-center replicated analysis. Pharmacoepidemiol Drug Saf. 2024 Mar;33(3):e5776. doi: 10.1002/pds.5776. PMID: 38479400.
 22. Goetze, S., van Drogen, A., Albinus, J.B. et al. Simultaneous targeted and discovery-driven clinical proteotyping using hybrid-PRM/DIA. *Clin Proteom* 21, 26 (2024). <https://doi.org/10.1186/s12014-024-09478-5>.
 23. Hall L, Dawel A, Greenwood LM, Monaghan C, Berryman K, Jack BN. Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components. Psychophysiology. 2023 Nov;60(11):e14363. doi: 10.1111/psyp.14363. Epub 2023 Jun 29. PMID: 37382363.
 24. Haven TL, Abunijela S, Hildebrand N. Biomedical supervisors' role modeling of open science practices. Elife. 2023 May 22;12:e83484. doi: 10.7554/eLife.83484. PMID: 37211820; PMCID: PMC10202448.
 25. Hayashi S, Caron BA, Heinsfeld AS, Vinci-Booher S, McPherson B, Bullock DN, Bertò G, Niso G, Hanekamp S, Levitas D, Ray K, MacKenzie A, Avesani P, Kitchell L, Leong JK, Nascimento-Silva F, Koudoro S, Willis H, Jolly JK, Pisner D, Zuidema TR, Kurzawski JW, Mikellidou K, Bussalab A, Chaumon M, George N, Rorden C, Victory C, Bhatia D, Aydogan DB, Yeh FF, Delogu F, Guaje J, Veraart J, Fischer J, Faskowitz J, Fabrega R, Hunt D, McKee S, Brown ST, Heyman S, Iacovella V, Mejia AF, Marinazzo D, Craddock RC, Olivetti E, Hanson JL, Garyfallidis E, Stanzione D, Carson J, Henschel R, Hancock DY, Stewart CA, Schnyer D, Eke DO, Poldrack RA, Bollmann S, Stewart A, Bridge H, Sani I, Freiwald WA, Puce A, Port NL, Pestilli F. brainlife.io:

- a decentralized and open-source cloud platform to support neuroscience research. *Nat Methods*. 2024 May;21(5):809-813. doi: 10.1038/s41592-024-02237-2. Epub 2024 Apr 11. Erratum in: *Nat Methods*. 2024 Jun;21(6):1131. doi: 10.1038/s41592-024-02296-5. PMID: 38605111; PMCID: PMC11093740.
26. Hellen DJ, Karpen SJ. *LiverQuant*: An Improved Method for Quantitative Analysis of Liver Pathology. *Bio Protoc*. 2023 Jul 20;13(14):e4776. doi: 10.21769/BioProtoc.4776. PMID: 37497459; PMCID: PMC10367012.
 27. Huckvale ED, Powell CD, Jin H, Moseley HNB. Benchmark Dataset for Training Machine Learning Models to Predict the Pathway Involvement of Metabolites. *Metabolites*. 2023 Nov 1;13(11):1120. doi: 10.3390/metabo13111120. PMID: 37999216; PMCID: PMC10673125.
 28. Jaric, I., Voelkl, B., Amrein, I. *et al*. Using mice from different breeding sites fails to improve replicability of results from single-laboratory studies. *Lab Anim* **53**, 18–22 (2024). <https://doi.org/10.1038/s41684-023-01307-w>.
 29. Kalvaitytė M, Gabrilavičiūtė S, Balciunas D. Rapid generation of single-insertion transgenics by Tol2 transposition in zebrafish. *Dev Dyn*. 2024 Nov;253(11):1056-1065. doi: 10.1002/dvdy.719. Epub 2024 Jun 30. PMID: 38946125.
 30. Kang, Kaidi & Seidlitz, Jakob & Bethlehem, Richard A.I. & Schildcrout, Jonathan & Tao, Ran & Xiong, Jiangmei & Jones, Megan & Alexander-Bloch, Aaron & Vandekar, Simon. (2023). Study design features that improve effect sizes in cross-sectional and longitudinal brain-wide association studies. *bioRxiv* : the preprint server for biology. 10.1101/2023.05.29.542742.
 31. Kekecs Z, Palfi B, Szaszi B, Szecsi P, Zrubka M, Kovacs M, Bakos BE, Cousineau D, Tressoldi P, Schmidt K, Grassi M, Evans TR, Yamada Y, Miller JK, Liu H, Yonemitsu F, Dubrov D, Röer JP, Becker M, Schnepfer R, Ariga A, Arriaga P, Oliveira R, Pöldver N, Kreegipuu K, Hall B, Wiechert S, Verschuere B, Girán K, Aczel B. Raising the value of research studies in psychological science by increasing the credibility of research reports: the transparent Psi project. *R Soc Open Sci*. 2023 Feb 1;10(2):191375. doi: 10.1098/rsos.191375. Erratum in: *R Soc Open Sci*. 2023 Aug 16;10(8):231080. doi: 10.1098/rsos.231080. PMID: 36756055; PMCID: PMC9890107.
 32. Kessel, Marcus & Atkinson, Colin. (2024). Promoting open science in test-driven software experiments. *Journal of Systems and Software*. 212. 111971. 10.1016/j.jss.2024.111971.
 33. Kolokathis, Panagiotis & Sidiropoulos, Nikolas & Zouraris, Dimitrios & Varsou, Dimitra-Danai & Mintis, Dimitris & Tsoumanis, Andreas & Dondero, Francesco & Exner, Thomas & Sarimveis, Haralambos & Chaideftou, Eugenia & Paparella, Martin & Nikiforou, Fotini & Karakoltzidis, Achilleas & Karakitsios, Spyros & Sarigiannis, Dimosthenis & Friis, Jesper & Goldbeck, Gerhard & Winkler, David & Peijnenburg, Willie & Afantitis, Antreas. (2024). Easy-MODA: Simplifying Standardised Registration of Scientific Simulation Workflows through MODA Template Guidelines powered by the Enalos Cloud Platform. *Computational and Structural Biotechnology Journal*. 25. 10.1016/j.csbj.2024.10.018.
 34. Krämer D, Schächtele L, Schneck A. Care to share? Experimental evidence on code sharing behavior in the social sciences. *PLoS One*. 2023 Aug 7;18(8):e0289380. doi: 10.1371/journal.pone.0289380. PMID: 37549146; PMCID: PMC10406284.
 35. Landau SM, Harrison TM, Baker SL, Boswell MS, Lee J, Taggett J, Ward TJ, Chadwick T, Murphy A, DeCarli C, Schwarz CG, Vemuri P, Jack CR Jr, Koeppe RA, Jagust WJ; U.S. POINTER Study Group and for the Alzheimer's Disease Neuroimaging Initiative. Positron emission tomography harmonization in the Alzheimer's Disease Neuroimaging Initiative: A scalable and rigorous approach to multisite amyloid and tau quantification. *Alzheimers Dement*. 2025 Jan;21(1):e14378. doi: 10.1002/alz.14378. Epub 2024 Nov 19. PMID: 39559932; PMCID: PMC11772732.
 36. Li X, Liu J, Gao W, Cohen GL. Challenging the N-Heuristic: Effect size, not sample size, predicts the replicability of psychological science. *PLoS One*. 2024 Aug 23;19(8):e0306911. doi: 10.1371/journal.pone.0306911. PMID: 39178270; PMCID: PMC11343368.
 37. Liu Q, Hu Q, Liu S, Hutson A, Morgan M. ReUseData: an R/Bioconductor tool for reusable and reproducible genomic data management. *BMC Bioinformatics*. 2024 Jan 3;25(1):8. doi: 10.1186/s12859-023-05626-0. PMID: 38172657; PMCID: PMC10765726.

38. Llovera G, Langhauser F, Isla Cainzos S, Hoppen M, Abberger H, Mohamud Yusuf A, Mencl S, Heindl S, Ricci A, Haupteltshofer S, Kuchenbecker-Pöls L, Gunzer M, Hansen W, Hermann DM, Gelderblom M, Schmidt-Pogoda A, Minnerup J, Kleinschnitz C, Magnus T, Liesz A. Stroke of Consistency: Streamlining Multicenter Protocols for Enhanced Reproducibility of Infarct Volumes in Preclinical Stroke Research. *Stroke*. 2024 Oct;55(10):2522-2527. doi: 10.1161/STROKEAHA.124.047232. Epub 2024 Sep 24. PMID: 39315830.
39. Luna Santamaría, M., Andersson, D., Parris, T.Z. *et al.* Digital RNA sequencing using unique molecular identifiers enables ultrasensitive RNA mutation analysis. *Commun Biol* 7, 249 (2024). <https://doi.org/10.1038/s42003-024-05955-7>.
40. Lusa L, Proust-Lima C, Schmidt CO, Lee KJ, le Cessie S, Baillie M, Lawrence F, Huebner M; TG3 of the STRATOS Initiative. Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis. *PLoS One*. 2024 May 29;19(5):e0295726. doi: 10.1371/journal.pone.0295726. PMID: 38809844; PMCID: PMC11135704.
41. Mager M, Pineda Hernandez H, Brandenburg F, López-Maury L, McCormick AJ, Nürnberg DJ, Orthwein T, Russo DA, Victoria AJ, Wang X, Zedler JAZ, Branco Dos Santos F, Schmelling NM. Interlaboratory Reproducibility in Growth and Reporter Expression in the Cyanobacterium *Synechocystis* sp. PCC 6803. *ACS Synth Biol*. 2023 Jun 16;12(6):1823-1835. doi: 10.1021/acssynbio.3c00150. Epub 2023 May 29. PMID: 37246820; PMCID: PMC10278186.
42. Martinc, M., Perovšek, M., Lavrač, N. *et al.* Textflows: an open science NLP evaluation approach. *Lang Resources & Evaluation* 59, 4439–4468 (2025). <https://doi.org/10.1007/s10579-024-09793-1>.
43. Martin D, Basodi S, Panta S, Rootes-Murdy K, Prae P, Sarwate AD, Kelly R, Romero J, Baker BT, Gazula H, Bockholt J, Turner JA, Esper NB, Franco AR, Plis S, Calhoun VD. Enhancing collaborative neuroimaging research: introducing COINSTAC Vaults for federated analysis and reproducibility. *Front Neuroinform*. 2023 Jun 19;17:1207721. doi: 10.3389/fninf.2023.1207721. PMID: 37404336; PMCID: PMC10315678.
44. Miyazaki T, Mizoguchi R, Ueda K, Shinozaki T, Kamoto M, Takeda Y, Sakuma S, Ito N, Momo M, Kawakami K. Crystallization of Amorphous Nifedipine Under Isothermal Conditions: Inter-laboratory Reproducibility and Investigation of the Factors Affecting Reproducibility. *J Pharm Sci*. 2023 Oct;112(10):2703-2716. doi: 10.1016/j.xphs.2023.06.002. Epub 2023 Jun 9. PMID: 37301322.
45. Nance N, Petersen ML, van der Laan M, Balzer LB. The Causal Roadmap and Simulations to Improve the Rigor and Reproducibility of Real-data Applications. *Epidemiology*. 2024 Nov 1;35(6):791-800. doi: 10.1097/EDE.0000000000001773. Epub 2024 Aug 1. PMID: 39087681; PMCID: PMC11444352.
46. Omondi C, Chou A, Fond KA, Morioka K, Joseph NR, Sacramento JA, Iorio E, Torres-Espin A, Radabaugh HL, Davis JA, Gumbel JH, Huie JR, Ferguson AR. Improving rigor and reproducibility in western blot experiments with the blotRig analysis. *Sci Rep*. 2024 Sep 17;14(1):21644. doi: 10.1038/s41598-024-70096-0. PMID: 39284854; PMCID: PMC11405887.
47. Payares-Garcia, David & Mateu, Jorge & Schick, Wiebke. (2023). Neuronorm: An R Package to Standardize Multiple Structural MRI. SSRN Electronic Journal. 10.2139/ssrn.4374278.
48. Pelletier SJ, Leclercq M, Roux-Dalvai F, de Geus MB, Leslie S, Wang W, Lam TT, Nairn AC, Arnold SE, Carlyle BC, Precioso F, Droit A. BERNN: Enhancing classification of Liquid Chromatography Mass Spectrometry data with batch effect removal neural networks. *Nat Commun*. 2024 May 6;15(1):3777. doi: 10.1038/s41467-024-48177-5. PMID: 38710683; PMCID: PMC11074280.
49. Rehal S, Cro S, Phillips PP, Fielding K, Carpenter JR. Handling intercurrent events and missing data in non-inferiority trials using the estimand framework: A tuberculosis case study. *Clin Trials*. 2023 Oct;20(5):497-506. doi: 10.1177/17407745231176773. Epub 2023 Jun 5. PMID: 37277978; PMCID: PMC10504812.
50. Rengasamy, M., Moriarity, D., Kraynak, T. *et al.* Exploring the multiverse: the impact of researchers' analytic decisions on relationships between depression and inflammatory markers. *Neuropsychopharmacol*. 48, 1465–1474 (2023). <https://doi.org/10.1038/s41386-023-01621-4>.
51. Rieger J, Jentsch C, Rahnenführer J. LDAPrototype: a model selection algorithm to improve reliability of latent Dirichlet allocation. *PeerJ Comput Sci*. 2024 Sep 20;10:e2279. doi: 10.7717/peerj-cs.2279. PMID: 39678270; PMCID: PMC11639148.

52. Rojas-Velazquez D, Kidwai S, Kraneveld AD, Tonda A, Oberski D, Garssen J, Lopez-Rincon A. Methodology for biomarker discovery with reproducibility in microbiome data using machine learning. *BMC Bioinformatics*. 2024 Jan 15;25(1):26. doi: 10.1186/s12859-024-05639-3. PMID: 38225565; PMCID: PMC10789030.
53. Schacherer DP, Herrmann MD, Clunie DA, Höfener H, Clifford W, Longabaugh WJR, Pieper S, Kikinis R, Fedorov A, Homeyer A. The NCI Imaging Data Commons as a platform for reproducible research in computational pathology. *Comput Methods Programs Biomed*. 2023 Dec;242:107839. doi: 10.1016/j.cmpb.2023.107839. Epub 2023 Oct 2. PMID: 37832430; PMCID: PMC10841477.
54. Schuetz I, Karimpur H, Fiehler K. vexptoolbox: A software toolbox for human behavior studies using the Vizard virtual reality platform. *Behav Res Methods*. 2023 Feb;55(2):570-582. doi: 10.3758/s13428-022-01831-6. Epub 2022 Mar 23. PMID: 35322350; PMCID: PMC10027796.
55. Senior AM. Estimating Genetic Variance in Life-Span Response to Diet: Insights From Statistical Simulation. *J Gerontol A Biol Sci Med Sci*. 2023 Mar 1;78(3):392-396. doi: 10.1093/gerona/glac172. PMID: 36008106; PMCID: PMC9977232.
56. Souza Junior DR, Silva ARM, Ronsein GE. Strategies for consistent and automated quantification of HDL proteome using data-independent acquisition. *J Lipid Res*. 2023 Jul;64(7):100397. doi: 10.1016/j.jlr.2023.100397. Epub 2023 Jun 5. PMID: 37286042; PMCID: PMC10339053.
57. Sprang M, Möllmann J, Andrade-Navarro MA, Fontaine JF. Overlooked poor-quality patient samples in sequencing data impair reproducibility of published clinically relevant datasets. *Genome Biol*. 2024 Aug 16;25(1):222. doi: 10.1186/s13059-024-03331-6. PMID: 39152483; PMCID: PMC11328481.
58. Stahl AC, Tietz AS, Kendziora B, Dewey M. Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in European Radiology? *Eur Radiol*. 2023 Jan;33(1):97-105. doi: 10.1007/s00330-022-09008-7. Epub 2022 Jul 30. PMID: 35907025; PMCID: PMC9362582.
59. Tang C, Huang D, Xing X, Yang H. EigenRF: an improved metabolomics normalization method with scores for reproducibility evaluation on importance rankings of differential metabolites. *Anal Methods*. 2024 Dec 19;17(1):45-53. doi: 10.1039/d4ay01569j. PMID: 39560372.
60. Tång Hallbäck E, Björkman JT, Dyrkell F, Welander J, Fang H, Sylvén I, Kaden R, Eilers H, Söderlund Strand A, Mernelius S, Berglind L, Campillay Lagos A, Engstrand L, Sikora P, Mölling P. Evaluation of nationwide analysis surveillance for methicillin-resistant *Staphylococcus aureus* within Genomic Medicine Sweden. *Microb Genom*. 2025 Jan;11(1):001331. doi: 10.1099/mgen.0.001331. PMID: 39869391; PMCID: PMC11893271.
61. van Houdt PJ, Ragunathan S, Berks M, Ahmed Z, Kershaw LE, Gurney-Champion OJ, Tadimalla S, Arvidsson J, Sun Y, Kallehauge J, Dickie B, Lévy S, Bell L, Sourbron S, Thrippleton MJ; on behalf of the ISMRM Perfusion Study Group. Contrast-agent-based perfusion MRI code repository and testing framework: ISMRM Open Science Initiative for Perfusion Imaging (OSIPI). *Magn Reson Med*. 2024 May;91(5):1774-1786. doi: 10.1002/mrm.29826. Epub 2023 Sep 4. PMID: 37667526.
62. van Kampen AHC, Mahamune U, Jongejan A, van Schaik BDC, Balashova D, Lashgari D, Pras-Raves M, Wever EJM, Dane AD, García-Valiente R, Moerland PD. ENCORE: a practical implementation to improve reproducibility and transparency of computational research. *Nat Commun*. 2024 Sep 16;15(1):8117. doi: 10.1038/s41467-024-52446-8. PMID: 39284801; PMCID: PMC11405857.
63. Viant MR, Amstalden E, Athersuch T, Bouhifd M, Camuzeaux S, Crizer DM, Driemert P, Ebbels T, Ekman D, Flick B, Giri V, Gómez-Romero M, Haake V, Herold M, Kende A, Lai F, Leonards PEG, Lim PP, Lloyd GR, Mosley J, Namini C, Rice JR, Romano S, Sands C, Smith MJ, Sobanski T, Southam AD, Swindale L, van Ravenzwaay B, Walk T, Weber RJM, Zickgraf FM, Kamp H. Demonstrating the reliability of in vivo metabolomics based chemical grouping: towards best practice. *Arch Toxicol*. 2024 Apr;98(4):1111-1123. doi: 10.1007/s00204-024-03680-y. Epub 2024 Feb 18. PMID: 38368582; PMCID: PMC10944399.
64. Wallace ML, Redline S, Oryshkewych N, Hoepel SJW, Luik AI, Stone KL, Kolko RP, Chung J, Leng Y, Robbins R, Zhang Y, Barnes LL, Lim AS, Yu L, Buysse DJ. Pioneering a multi-phase framework to harmonize self-reported sleep data across cohorts. *Sleep*. 2024 Sep 9;47(9):zsae115. doi: 10.1093/sleep/zsae115. PMID: 38752786; PMCID: PMC11381567.

65. Wirz, C. D., Sutter, C., Demuth, J. L., Mayer, K., Chapman, W., Cains, M., Radford, J., Przybylo, V., Evans, A., Martin, T., Gaudet, L. C., Sulia, K., Bostrom, A., Gagne, D. J., Bassill, N., Schumacher, A., & Thorncroft, C. (2024). Increasing the reproducibility and replicability of supervised AI/ML in the Earth systems science by leveraging social science methods. *Earth and Space Science*, 11, Article e2023EA003364. <https://doi.org/10.1029/2023EA003364> (Original work published 2024).
66. Yousefirizi F, Liyanage A, Klyuzhin IS, Rahmim A. From code sharing to sharing of implementations: Advancing reproducible AI development for medical imaging through federated testing. *J Med Imaging Radiat Sci*. 2024 Dec;55(4):101745. doi: 10.1016/j.jmir.2024.101745. Epub 2024 Aug 29. PMID: 39208523.
67. Zhang XS, Xie G, Ma H, Ding S, Wu YX, Fei Y, Cheng Q, Huang Y, Wang Y. Highly reproducible and cost-effective one-pot organoid differentiation using a novel platform based on PF-127 triggered spheroid assembly. *Biofabrication*. 2023 Aug 21;15(4). doi: 10.1088/1758-5090/acee21. PMID: 37552975.
68. Zhiyi Zhu, Min Chen, Zhen Qian, Hengyue Li, Kai Wu, Zaiyang Ma, Yongning Wen, Songshan Yue, and Guonian Lü. 2023. Documentation strategy for facilitating the reproducibility of geo-simulation experiments. *Environ. Model. Softw.* 163, C (May 2023). <https://doi.org/10.1016/j.envsoft.2023.105687>.

5.4 Data extraction items and their answering options

1. Source, options:

- 1.1. Peer-reviewed
- 1.2. Pre-print
- 1.3. Grey literature

2. Study design category, options

- 2.1. Comparative (between-subject comparison)
- 2.2. Comparative (within-subject comparison/repeated measures design)
- 2.3. Post-test design (only a post measurement after the implementation of an intervention and the intervention is explicitly mentioned)
- 2.4. Qualitative
- 2.5. Other
- 2.6. Not applicable
- 2.7. Not reported

3. Broad academic field classification, options:

- 3.1. Natural sciences
- 3.2. Engineering and technology
- 3.3. Medical and health sciences
- 3.4. Agricultural and veterinary sciences
- 3.5. Social sciences
- 3.6. Humanities and the arts
- 3.7. Not applicable
- 3.8. Not reported

4. Sample size; open text field

5. Intervention classification, (multiple) options:

- 5.1. Open methodology
- 5.2. Open science policies and guidelines
- 5.3. Open data and materials
- 5.4. Open science ethics
- 5.5. Open evaluations
- 5.6. Open educational resources
- 5.7. Open science tools
- 5.8. Other

6. Intervention Details

- 6.1. Level of independent variable (Who installs/applies/implements the intervention), options:
 - 6.6.1. Government
 - 6.6.2. Funder
 - 6.6.3. Publisher
 - 6.6.4. Journal
 - 6.6.5. Institution
 - 6.6.6. Reviewers
 - 6.6.7. PhD students
 - 6.6.8. Post docs

- 6.6.9. Academic staff
- 6.6.10. Academic support staff
- 6.6.11. Other
- 6.6.12. Not applicable
- 6.2. Level of dependent variable (Who is affected), options:
 - 6.2.1. Government
 - 6.2.2. Funder
 - 6.2.3. Publisher
 - 6.2.4. Journal
 - 6.2.5. Institution
 - 6.2.6. Reviewers
 - 6.2.7. PhD students
 - 6.2.8. Post docs
 - 6.2.9. Academic staff
 - 6.2.10. Academic support staff
 - 6.2.11. Other
 - 6.2.12. Not applicable
- 6.3. Stage of the process, options:
 - 6.3.1. Before research
 - 6.3.2. During research
 - 6.3.3. After research before or at publication
 - 6.3.4. After publication
 - 6.3.5. Other

7. Outcome, options:

- 7.1. Direct outcomes, options:
- 7.2. Inferential reproducibility
- 7.3. Computational/Outcome/Result/Numerical/Statistical reproducibility
- 7.4. Method /methodological reproducibility
- 7.5. Inferential replicability
- 7.6. Computational/Outcome/Result/Numerical/Statistical replicability
- 7.7. Other

8. Proxy outcomes, multiple options:

- 8.1. Registration Status
- 8.2. Review
- 8.3. Methods transparency
- 8.4. Data
- 8.5. Research material sharing

9. Important limitations: tick box (Y/N) plus open text field

10. Exclude Out of scope: tick box (Y/N)

11. Data extraction finished: tick box (Y/N)

A table that presents the data extraction items of the TIER2/OSIRIS scoping review that were excluded from the TRUSTparenacy data extraction:

Design	Sample	Sample description	Intervention details	Outcome	Results

Research Question-Hypothesis (Open text field) Experimental design characteristics (options: Manipulation, Randomization, Not applicable, Not reported)	Sample (options: Articles/studies, PhD students, Post docs, Academic staff, Reviewers, Academic support staff, Journals, Publishers, Institutions, Funders, Other)	Country (options: Country - open text field, Not applicable, Not reported) Broad academic field classification (options: Other)	Stringency (options: Mandatory, Optional, Other, Not applicable, Not reported)	Outcome Details (options: Outcome metric-operationalization, When was outcome measured, in relation to intervention)	Results (options: Answer to Research Question/Hypothesis, Relevant descriptives, Primary statistical-meta analytic or qualitative results for each hypothesis, Additional results.
--	---	--	--	--	---

Under review by the European Commission

5.5 Data extracted (1/3) – Context of the intervention

No.	First author surname	Year	Academic field	Who implements the intervention	Who is affected by the intervention	At which stage of the process
1	Abdelaziz	2023	Medical and health sciences	PhD students	Researchers	During research
2	Aguilar Gomez	2023	Engineering and technology	Researchers	Researchers	During research
3	Ahern	2024	Medical and health sciences	Researchers	Researchers	After publication
4	Anido-Alonso	2023	Medical and health sciences	Researchers	Researchers	During research
5	Ashkarran	2024	Medical and health sciences	Researchers	Researchers	During research
6	Bahl	2023	Medical and health sciences	Researchers	Researchers	After research before or at publication
7	Bontempi	2024	Medical and health sciences	Academic staff	Researchers	During research
8	Bramley	2024	Medical and health sciences	Journal	Journal	After research before or at publication
9	Chen	2025	Medical and health sciences	Researchers	Researchers	During research
10	Denisova	2024	Engineering and technology	Researchers	Researchers	During research
11	De Vries	2024	Natural sciences	Researchers	Researchers	During research
12	Ellis	2024	Social sciences	Researchers	Researchers	After research before or at publication
13	Elmanzalawi	2025	Natural sciences	Researchers	Researchers	During research
14	Fernholz	2024	Medical and health sciences	Researchers	Researchers	During research

15	Ferreira	2023	Agricultural and veterinary sciences	Researchers	Researchers	During research
16	Fiedler	2023	Agricultural and veterinary sciences	Researchers	Researchers	During research
17	Filatovas	2024	Engineering and technology	Researchers	Researchers	During research
18	Fraser	2023	Social sciences	Researchers	Researchers	After research before or at publication
19	Gaignard	2023	Medical and health sciences	Researchers	Researchers	After research before or at publication
20	Gierend	2023	Medical and health sciences	Researchers	Researchers	During research
21	Gillies	2024	Medical and health sciences	Researchers	Institution	During research
22	Goetze	2024	Medical and health sciences	Researchers	Researchers	During research
23	Hall	2023	Medical and health sciences	Researchers	Researchers	Before research
24	Haven	2023	Medical and health sciences	Academic staff	PhD students	After publication
25	Hayashi	2023	Medical and health sciences	Researchers	Researchers	During research
26	Hellen	2023	Medical and health sciences	Researchers	Researchers	During research
27	Huckvale	2023	Natural sciences	Researchers	Researchers	After research before or at publication
28	Jaric	2024	Agricultural and veterinary sciences	Researchers	Researchers	Before research
29	Kalvaityte	2024	Natural sciences	Researchers	Researchers	During research
30	Kang	2023	Medical and health sciences	Researchers	Researchers	During research

31	Kekecs	2023	Social sciences	Researchers	Researchers	Before research
32	Kessel	2024	Engineering and technology	Researchers	Researchers	During research
33	Kolokathis	2024	Engineering and technology	Researchers	Researchers	During research
34	Krahmer	2023	Social sciences	Researchers	Researchers	After research before or at publication
35	Landau	2024	Medical and health sciences	Researchers	Researchers	During research
36	Liu	2024	Social sciences	Researchers	Researchers	After research before or at publication
37	Li	2024	Medical and health sciences	Researchers	Researchers	During research
38	Llovera	2024	Medical and health sciences	Researchers	Researchers	Before research
39	Luna	2024	Medical and health sciences	Researchers	Researchers	During research
40	Lusa	2024	Not applicable	Researchers	Researchers	During research
41	Mager	2023	Natural sciences	Researchers	Researchers	During research
42	Martinc	2024	Humanities and the arts	Researchers	Researchers	After research before or at publication
43	Martin	2023	Medical and health sciences	Researchers	Researchers	After research before or at publication
44	Miyazaki	2023	Natural sciences	Researchers	Researchers	During research
45	Nance	2024	Not reported	Researchers	Researchers	During research
46	Omondi	2024	Natural sciences	Researchers	Researchers	During research
47	Payares-Garcia	2023	Engineering and technology	Researchers	Researchers	During research

48	Pelletier	2024	Natural sciences	Researchers	Researchers	During research
49	Rehal	2023	Medical and health sciences	Researchers	Researchers	After publication
50	Rengasamy	2023	Medical and health sciences	Researchers	Researchers	During research
51	Rieger	2024	Humanities and the arts	Researchers	Researchers	During research
52	Rojas-Velazquez	2024	Medical and health sciences	Researchers	Researchers	During research
53	Schacherer	2023	Medical and health sciences	Researchers	Researchers	During research
54	Schuetz	2023	Social sciences	Researchers	Researchers	During research
55	Senior	2023	Medical and health sciences	Researchers	Researchers	During research
56	Souza	2023	Medical and health sciences	Researchers	Researchers	After research before or at publication
57	Sprang	2024	Medical and health sciences	Researchers	Researchers	During research
58	Stahl	2023	Not applicable	Researchers	Researchers	After research before or at publication
59	Tang	2024	Medical and health sciences	Researchers	Researchers	After research before or at publication
60	Tång-Hallbäck	2025	Medical and health sciences	Researchers	Researchers	During research
61	van Houdt	2023	Engineering and technology	Researchers	Researchers	During research
62	Van Kampen	2024	Not applicable	Researchers	Researchers	During research
63	Viant	2024	Natural sciences	Researchers	Researchers	During research
64	Wallace	2024	Medical and health sciences	Researchers	Researchers	During research

65	Wirz	2024	Engineering and technology	Researchers	Researchers	During research
66	Yousefirizi	2024	Engineering and technology	Researchers	Researchers	During research
67	Zhang	2023	Medical and health sciences	Researchers	Researchers	During research
68	Zhu	2023	Natural sciences	Researchers	Researchers	During research

5.6 Data extracted (2/3) – Characteristics of the intervention

No.	First author surname	Year	Design	Sample size	Intervention classification	Intervention sub-classification	Outcome	Proxy outcome class	Proxy outcome sub-class
1	Abdelaziz	2023	Comparative (within-subject comparison/repeated measures design)	not explicitly stated (Size range from 10%-43% of original data)	Open methodology	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
2	Aguilar Gomez	2023	Comparative (between-subject comparison)	not explicitly stated	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	methods transparency general
3	Ahern	2024	Comparative (within-subject comparison/repeated measures design)	meta-analysis of 97 studies	Open data and materials	data availability	Method /methodological reproducibility	Methods transparency	analysis transparency
4	Anido-Alonso	2023	Comparative (between-subject comparison)	Six publicly available datasets	Open data and materials	data availability	Method /methodological reproducibility	Methods transparency	methods transparency general
5	Ashkarran	2024	Comparative (within-subject comparison/repeated measures design)	Not clearly stated	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
6	Bahl	2023	Comparative (within-subject comparison/repeated measures design)	meta-analysis of 25 proteomic datasets.	Open methodology	project workflow	Method /methodological reproducibility	Methods transparency	analysis transparency

7	Bontempi	2024	Comparative (within-subject comparison/repeated measures design)	Two use cases: Use case 1: Original subset N=206, extended dataset=421. Use case 2: Radiomics collection N=420, Radiogenomics collection N=133	Open science tools	open workflow tools, workflow, management systems	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	research process or workflow transparency
8	Bramley	2024	Post-test design (only a post measurement after the implementation of an intervention and the intervention is explicitly mentioned)	68 manuscripts	Open data and materials	data availability	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
9	Chen	2025	Comparative (within-subject comparison/repeated measures design)	Clinical samples not quantified	Open methodology	open protocol	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	methods transparency general
10	Denisova	2024	Post-test design (only a post measurement after the implementation of an intervention and the intervention is explicitly mentioned)	12 (6 clinicians and 6 engineers)	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
11	DeVries	2024	Comparative (within-subject comparison/repeated measures design)	3 case studies	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	analysis transparency
12	Ellis	2024	Comparative (between-subject comparison)	204 papers	Open science policies and guidelines	policy guidelines (e.g. of funders/publishers)	Method /methodological reproducibility	Methods transparency	methods transparency general
13	Elmanzalawi	2025	Comparative (within-subject comparison/repeated measures design)	5184-10000 genomes	Open data and materials	data availability	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	methods transparency general

14	Fernholz	2024	Comparative (between-subject comparison)	12 datasets	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	design transparency
15	Ferreira	2023	Comparative (between-subject comparison)	33 samples (21 flower samples, 12 genotype comparison)	Open methodology	project workflow	Method /methodological reproducibility	Methods transparency	methods transparency general
16	Fiedler	2023	Comparative (between-subject comparison)	Not applicable (protocol development, not hypothesis testing)	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
17	Filatovas	2024	Other	Not stated	Open science tools	open workflow tools, workflow, management systems	Other	Research material sharing	research material sharing general
18	Fraser	2023	Comparative (within-subject comparison/repeated measures design)	5 groups of 5 participants assessing 25 claims	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
19	Gaignard	2023	Post-test design (only a post measurement after the implementation of an intervention and the intervention is explicitly mentioned)	500 datasets each from Zenodo, Pangaea, and Dryad repositories	Open data and materials	data availability	Method /methodological reproducibility	Data	
20	Gierend	2023	Post-test design (only a post measurement after the implementation of an intervention and the intervention is explicitly mentioned)	Generated 100,000 data elements each for 7 data elements type to generate a total of 700,000 provenance records using a Python (Python Foundation) script.	Open data and materials	data availability	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
21	Gillies	2024	Other	53,127 metformin initiators	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	analysis transparency
22	Goetze	2024	Comparative (within-subject comparison/repeated measures design)	185 selected proteotypic peptides for tumor-associated antigens derived from	Open methodology	open protocol	Method /methodological reproducibility	Data	

				64 annotated human protein groups. Applied hybrid-PRM/DIA for the integrated digitization of biobanked melanoma samples using a set of 30 AQUA peptides against 28 biomarker candidates					
23	Hall	2023	Comparative (within-subject comparison/repeated measures design)	50 participants	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	methods transparency general
24	Haven	2023	Comparative (between-subject comparison)	211 pairs of supervisor and PhD candidates and 2062 publications	Open science policies and guidelines	on data/code sharing or open science practices	Method /methodological reproducibility	Data	
25	Hayashi	2023	Other	Not specified	Open science tools	open workflow tools, workflow, management systems	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	methods transparency general
26	Hellen	2023	Comparative (between-subject comparison)	Not specified	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	analysis transparency
27	Huckvale	2023	Other	Final dataset had 5683 metabolites	Open data and materials	data availability	Method /methodological reproducibility	Data	
28	Jaric	2024	Comparative (between-subject comparison)	308 mice	Other		Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	methods transparency general
29	Kalvaityte	2024	Comparative (between-subject comparison)	Not stated explicitly	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
30	Kang	2023	Comparative (within-subject comparison/repeated measures design)	77695 scans	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
31	Kekecs	2023	Other	2115 participants	Open methodology	pre-registration, study registration, analysis plan	Method /methodological reproducibility	Methods transparency	methods transparency general

32	Kessel	2024	Comparative (between-subject comparison)	94800 component test pairs	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
33	Kolokathis	2024	Other	Not applicable	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	methods transparency general
34	Krahmer	2023	Comparative (between-subject comparison)	1206 contacted authors	Open science policies and guidelines	on data/code sharing or open science practices	Method /methodological reproducibility	Methods transparency	methods transparency general
35	Landau	2024	Other	4280 participants	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
36	Liu	2024	Comparative (between-subject comparison)	307 for original replication study pairs 284 for effect size analysis due to missing data 215 for survey of psychologists	Open science policies and guidelines	reporting guidelines, reporting standards	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	design transparency
37	Li	2024	Other	Not applicable	Open science tools	open workflow tools, workflow, management systems	Method /methodological reproducibility	Methods transparency	methods transparency general
38	Llovera	2024	Comparative (between-subject comparison)	245 mice	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
39	Luna	2024	Comparative (between-subject comparison)	Not explicitly stated	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	methods transparency general
40	Lusa	2024	Other	One case study	Open science policies and guidelines	on reporting checklists or guidance	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
41	Mager	2023	Comparative (within-subject comparison/repeated measures design)	Participants from eight different laboratories quantified the fluorescence intensity of mVENUS as a proxy for the transcription activity	Open science policies and guidelines	on reporting checklists or guidance	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency

				of the three promoters.					
42	Martinc	2024	Qualitative	The authors evaluated and compared seven open-source tools for part-of-speech tagging on six different corpora, analyzing the effects of different factors that may affect the accuracy of the part-of-speech tagging tools, such as, for example, the training corpus length.	Open data and materials	data availability	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	research process or workflow transparency
43	Martin	2023	Other	Five use-cases that highlight the benefits and versatility of Vaults in COINSTAC.	Open data and materials	data availability	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
44	Miyazaki	2023	Comparative (within-subject comparison/repeated measures design)	The investigation to standardize the protocol was conducted by the six partner laboratories.	Open science policies and guidelines	on research quality	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	research process or workflow transparency
45	Nance	2024	Comparative (between-subject comparison)	We illustrate with two worked examples. First, in an observational longitudinal study, second, in a cluster randomized trial with missing outcomes.	Open methodology	project workflow	Inferential reproducibility	Methods transparency	analysis transparency
46	Omondi	2024	Comparative (within-subject comparison/repeated measures design)	The authors developed the blotRig protocol and application harnessing a database of 6000 + western blots from N = 281 subjects (rats and mice) collected by multiple UCSF labs on	Open science policies and guidelines	reporting guidelines, reporting standards	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency

				core equipment. To demonstrate blotRig best practices in a real-world experiment, we carried out prospective multiplexed WB analysis of protein lysate from lumbar cord in rodent models of spinal cord injury (SCI) (N = 29 rats) in 2 groups (experimental group & control group).					
47	Payares-Garcia	2023	Comparative (within-subject comparison/repeated measures design)	The authors tested NeuroNorm with MRI data of 330 patients from from three different data sources (the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Parkinson's Progression Markers Initiative (PPMI), and the Multiple Sclerosis database of the University Medical Center Ljubljana.).	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	research process or workflow transparency
48	Pelletier	2024	Comparative (within-subject comparison/repeated measures design)	We are using three datasets with various levels of batch mixing heterogeneity and different numbers of batches to demonstrate how BERNN can apply to different scenarios.	Open science policies and guidelines	on research quality	Method /methodological reproducibility	Methods transparency	analysis transparency
49	Rehal	2023	Other	3-arm double-blind, placebo-controlled non-inferiority trial,	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency

				randomising 1931 patients 1:1:1 to either the standard treatment regimen or two shorter treatment combinations, in patients with newly diagnosed pulmonary tuberculosis.			merical/Statistical reproducibility		
50	Rengasamy	2023	Other	Participants included in this analysis were 25,962 adults who provided data across any of five cycles of NHANES with valid PHQ scores and also either WBC or CRP levels.	Other		Inferential reproducibility	Methods transparency	analysis transparency
51	Rieger	2024	Comparative (between-subject comparison)	Six databases, with more than 5 million articles (see Table 2).	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
52	Rojas-Velazquez	2024	Comparative (between-subject comparison)	We used a total of nine datasets, with three datasets for each experiment.	Open science policies and guidelines	on reporting checklists or guidance	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
53	Schacherer	2023	Comparative (within-subject comparison/repeated measures design)	To investigate if and how this affects reproducibility, both experiments were executed multiple times, each in a new instance of the respective ML service.	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
54	Schuetz	2023	Qualitative	The authors implemented a simple VR experiment, followed by summarizing some of the behavioral data and measures from a small-scale (N = 5)	Open methodology	open protocol	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	

				example data collection performed in our lab.					
55	Senior	2023	Other	The author estimated the replicability of strain-specific effects of DR, i.e., strain-wise correlation of effects in 2 experiments.	Open science policies and guidelines	on research quality	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
56	Souza	2023	Comparative (within-subject comparison/repeated measures design)	We have chosen four freely available software that are both user friendly (present a graphical user interface) and require little to no special skills to analyze proteomics data. Plasma samples from six apparently healthy donors were obtained from blood collected in EDTA-containing vacuum tubes.	Open science policies and guidelines	on reporting checklists or guidance	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
57	Sprang	2024	Comparative (between-subject comparison)	The analysis uses 40 clinically relevant RNA-seq datasets. From the 40 datasets, our stringent manual selection included a total of 1164 human samples. There was an average number of samples equal to 29.1 per dataset ranging from 8 to 96.	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
58	Stahl	2023	Comparative (within-subject comparison/repeated measures design)	The database search on MEDLINE via PubMed identified 719 references. Another 657 studies were identified by manually searching	Open science policies and guidelines	reporting guidelines, reporting standards	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	

				the website of European Radiology.					
59	Tang	2024	Comparative (within-subject comparison/repeated measures design)	EigenRF was compared with EigenMS and ten other commonly used normalization methods across three publicly accessible datasets with distinct characteristics. This comparison serves to validate the generalizability of EigenRF.	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
60	Tång-Hallbäck	2025	Comparative (within-subject comparison/repeated measures design)	A national methicillin-resistant Staphylococcus aureus (MRSA) collection of 20 strains was distributed to nine participating laboratories that performed in-house procedures for WGS. Raw data were shared and analysed by three pipelines: 1928 Diagnostics, JASEN (GMS pipeline) and CLC-Genomics Workbench.	Open science policies and guidelines	reporting guidelines, reporting standards	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
61	van Houdt	2023	Comparative (within-subject comparison/repeated measures design)	At the time of writing, the taskforce received Python source code contributions to the repository, comprising 86 implementations of individual perfusion processing steps, contributed by 12 individuals or teams.	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	research process or workflow transparency

62	Van Kampen	2024	Not applicable	Nor mentioned	Open science policies and guidelines	on data/code sharing or open science practices	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
63	Viant	2024	Comparative (within-subject comparison/repeated measures design)	Six ringtrial partners worked independently on eight substances to prepare samples, acquire LC-MS metabolomics data, process and statistically analyze the data, and then report their chemical grouping results to ECHA.	Open science policies and guidelines	on research quality	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	analysis transparency
64	Wallace	2024	Qualitative	The authors produce a pooled multi-cohort sample of five US cohorts plus a separate yet fully harmonized sample from Rotterdam, Netherlands.	Open science policies and guidelines	reporting guidelines, reporting standards	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
65	Wirz	2024	Qualitative	For reliability trials they used approximately 50 images per trial, which satisfied the guideline described in Section 3.3.1 of about one focused hour of labeling.	Open methodology	project workflow	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	
66	Yousefirizi	2024	Comparative (within-subject comparison/repeated measures design)	Analysis of data from 36 PET/CT scans of patients (12 from Center I and 24 from the fully automated lesion segmentation in oncologic PET/CT imaging challenge dataset (autoPET)). They used 5 pre-	Open science tools	open workflow tools, workflow, management systems	Computational/Outcome/Result/Numerical/Statistical reproducibility	Data	

				processed cases from center II for the experiments related to the deployment environment and post-processing steps.					
67	Zhang	2023	Comparative (within-subject comparison/repeated measures design)	Development of 4 different types of organoids (hepatic, nephron, cortical, and lung). But no specifics on how many of them were developed.	Open methodology	open protocol	Method /methodological reproducibility	Methods transparency	research process or workflow transparency
68	Zhu	2023	Comparative (within-subject comparison/repeated measures design)	One case study.	Open science policies and guidelines	reporting guidelines, reporting standards	Computational/Outcome/Result/Numerical/Statistical reproducibility	Methods transparency	methods transparency general

5.7 Data extracted (3/3) – declared limitations of the study

No.	First author surname	Year	Limitations
1	Abdelaziz	2023	Not explicitly mentioned.
2	Aguilar Gomez	2023	- Although providing different weights to the level of importance of every single indicator (Essential, Important, Useful) is a good approach to emphasize some of the tests, some of them can be found repetitive, scoring more points for similar or almost equal checks. Also, as Table 3 shows, the distribution of Essential, Important and Useful indicators is somehow unbalanced, with the greatest share of Essential indicators under F and A (7 and 8, respectively), while one finds 7 such indicators under I and just 5 under R. This distribution has a clear influence on the assessment outcomes, and in a way one may argue that the ultimate goal of FAIR (that of maximize the reuse of data) gets compromised. - Furthermore, most of the “Essential” indicators are defined in F and A, which are those more closely related to the repository infrastructure (see Fig. 13). Therefore, depending on the data service, these tests will pass even if the user does not provide especially rich metadata or proper data, which also impact directly on the final score with significant weight. - Different tools return widely different results because of independent interpretations of the FAIR principles, indicators or metrics, as well as the data or metadata being gathered.
3	Ahern	2024	A limitation of our approach is that it addresses only confounding and random error; it assumes that no other systematic bias affects an evidence base. In the overweight-mortality evidence base, additional bias may exist in constituent studies from misclassification of adiposity by BMI, mismeasurement of BMI by self-report, and from selection bias.

			Another limitation is that variables in external datasets may not always map perfectly to the variables used in a study targeted for simulated re-adjustment. This concern warrants careful consideration on a case-by-case basis and may require that a different external dataset be chosen.
4	Anido-Alonso	2023	The performance of local models is influenced by the specifics of the source dataset, leading to poor inter-database generalization performance. In particular, while comparable performance between local validation and test partitions were observed, considerable performance drop was experienced when local models were targeted the prediction of data from external sources.
5	Ashkarran	2024	The difference (uniqueness) of the data from each core still outweighs their similarities, despite streamlining sample preparation and the instrumentation used.
6	Bahl	2023	- As it is often the case in repositories, the lack of metadata adjoining the datasets hampers proper comparison among treatments or conditions, as the relationship between raw files and corresponding measured samples is not clear. - The underlying assumption of MaxLFQ is that the majority of proteins is not changing between analyzed conditions. However, in the original publication of MaxLFQ, the authors tested this assumption in a benchmark dataset in which more than 30% of all identified proteins were changed. - The tool that we introduce in this work, PROTEOMAS (PROTEOmics Meta-AnalysiS), can perform automated and harmonized meta-analyses of data-dependent acquisition (DDA) proteomic datasets using the popular and commonly used label-free quantification (LFQ) algorithm of the freeware MaxQuant (suggesting a limitation in its direct application to other acquisition strategies without modification) - After assigning samples to their respective conditions, it is checked whether at least triplicates are present in each group as this is a minimal pre-requisite for successful outlier detection and statistical testing. If this is not the case for some of the conditions, those are deleted from the dataset.
7	Bontempi	2024	- Adoption challenges for users unaccustomed to cloud computing. Intellectual property rights can pose challenges, especially when such AI models are being validated (or planning to be validated) further in a clinical setting (e.g., in perspective clinical trials) or when they can be of commercial value for companies funding the research—as a significant part of AI research in radiology, and, more generally, medical imaging, is nowadays driven by industry. - Moreover, in such a highly competitive field, researchers are often caught in a race against time, striving for rapid development and deployment to maintain a competitive edge. This urgency can lead to a reluctance to fully embrace transparency, as research groups may fear that sharing too much information could have a negative impact on their advantage.
8	Bramley	2024	Fewer authors provided statistical code, which was a surprise given that this has fewer confidentiality implications and is straightforward in most statistical packages. This may be due to lack of technical expertise, which is supported by the fact that several groups provided documents labelled as code which were not statistical code. Despite this, more than half of submissions provided documents which could lead to full reproducibility of results – though I lacked the resources to evaluate whether all results were reproducible.
9	Chen	2025	The sponsor's lack of direct laboratory presence at the CROs limits their ability to effectively oversee these critical processes. The data in this manuscript represented ELISA methods, which are the common types of ligand binding assays used to measure biologic molecules.
10	Denisova	2024	- However, at the moment of preparing the first iteration of the usability testing procedure, the CT modality was the one with the highest level of readiness, leading us to limit the tasks in the administered tests, as described in this work, to CT acquisitions only. - It is worth noting, that the test description and the terminology used can greatly influence the success (or failure) of the test results. It is indeed very important to carefully formulate the tests and the tasks, asking, when possible, for the help of an external specialised end user for implementing the appropriate terminology and lexicon. - Remote testing may allow a larger group of participants while obtaining similar results as those performed in-person, but it shows some limitations and drawbacks. The main challenge of remote testing consists of the participants being in charge of the whole environment and equipment set-up, which may lead to an unwanted and unexpected set of problems.

			Another issue which might be faced are the possible different time zones for the participants and the test conductors.
11	DeVries	2024	Not explicitly mentioned.
12	Ellis	2024	- Given publication lags, it is difficult to completely ensure that all of the papers treated as CPAs were actually submitted after the policy changes occurred. For example, a 2021-published paper could have been accepted before the policy changes were enacted. This information was not available to us, but our decline analysis, which involved recording the number of days since publication for each paper, reveals minimal overlap. However, it is important to note that this decline analysis was correlational and would require a longer-term follow-up before stronger claims regarding cause and effect could be made. It remains possible that articles published after the policy changes will, over time, end up with an identical level of decay to that of earlier articles. - While not all significant, each measure showed some improvement following changes to journal policy. However, these may have improved as a result of coders becoming better at accessing resources as they worked through their sample (practice effects). Papers were selected at random, and coders could work through them in any order, but lists were delivered to them such that the PPAs appeared before the CPAs. - We note that primary asset type was not balanced between pairs and while the split was broadly consistent with that represented in Fig. 1, some pairs, for example, assessed more statistical code or materials than data from surveys. While unlikely, especially given an N of 34 per pair and a large variability between assets, it is difficult to determine categorically whether any of this variation has impacted our results. - The number of minutes needed to access and get research assets up and running was less useful than anticipated. Sometimes it was simply not possible to access any assets and so the clock never started (N/A was recorded). Similarly, N/A was also recorded when assets appeared on first inspection to be available, only for coders to immediately hit a brick wall thereafter. Where a time was recorded, it could naturally be very short if assets were complete and reusable. However, more complex and well-documented assets that were equally complete and reusable could still take considerable time to set up. - Our focus was on quantitative research assets and, based on previous research, our measures were also quantitative in nature (Roche et al., 2015; Towse et al., 2021a). This perhaps disguises a somewhat more complex landscape. Put simply, even with a highly experienced team, accessing a variety of research assets was far from straightforward, which is difficult to completely capture with three quantitative measures. - If we were experts in the fields of work for all papers, and if we were experienced in handling all manner of compilers, interpreters, and their troubleshooting and debugging, then the speed and overall success of the checks might have improved. And yet, if that meant we only assessed assets that had been checked by deeply knowledgeable others (where we were, in effect, insiders for the work), this would surely produce a biased estimate of useability.
13	Elmanzalawi	2025	Not explicitly mentioned.
14	Fernholz	2024	- First, our results call attention to the lack of reproducibility of studies involving spine quantification, since manual annotations are the current gold standard of spine counting and localization. - Furthermore, our findings emphasize the necessity to involve multiple annotators when evaluating a dataset. The manual work of annotators already represents the main bottleneck in spine analysis pipelines, and the apparent need to perform this task multiple times further decreases the already low throughput of this process. Most automated methods of spine quantification have been using manual spine annotations as ground truth for both training and validation of the method. Here, too, the observed amount of inter- and intra-rater reliability suggests that multiple annotators are required for training and validation of automated means of spine quantification in order to minimize the subjective bias a single user introduces.
15	Ferreira	2023	Not explicitly mentioned.
16	Fiedler	2023	Insufficient reproducibility and comparability of analytical results in ecotoxicology studies may result from underreporting of a study as well as from various confounding variables, such as different exposure concentrations and -systems, different ages, sexes or genetic background of the examined fish, or differing technical procedures applied in necropsy and sample processing.

17	Filatovas	2024	Not explicitly mentioned.
18	Fraser	2023	- A problem for any research involving human subjects is elicitation burden—both the quantity and quality of information provided by experts are thought to decline the longer the elicitation process takes, similar to participant fatigue in surveys and experiments. There is thus a trade-off between the extent of information desired for research purposes and the number and richness of questions asked. - A related difficulty is that research participants might have multiple interpretations of key terms for the elicitation. Indeed, there can be considerable conceptual slippage around many of these terms, such as conflicting taxonomies attempting to define replication practices. - Another challenge is that participants may hold opinions about aspects of a study that could bias their judgements about replicability. Recruitment challenges: The repliCATS recruitment strategy meant that we were not able fully to control for or consistently recruit diversity in groups.
19	Gaignard	2023	- Due to the genericity of FAIR principles and their nontechnical specifications, their implementation is highly dependent on human interpretation. - The A2 principle, addressing long-term preservation, is clearly out of the scope of Semantic Web technologies. F3 is hard to assess, since there is no consensus on the appropriate vocabulary nor on the associated cardinality constraints aimed at unambiguously identifying data.
20	Gierend	2023	- The library has only been tested with simulated data, not in a real environment. - Provenance alone cannot ensure data correctness and security against malicious users. - Needs further scalability analysis to expand the current provenance class.
21	Gillies	2024	- A limitation of our study was that our defined thresholds for our concordance ratings were consensus-based and do not necessarily reflect clinically important differences. - Further, PBS data only include subsidized medicine dispensings and therefore some diabetes medicine use was overlooked in our analysis; this limitation would not have affected our concordance results. - Additionally, though each participating site had ethical and data custodian approval to access the PBS 10% sample data, patient-level data could not be shared between sites and this restriction meant we were unable to investigate differences at the individual level between sites.
22	Goetze	2024	One of the challenges of the hybrid-PRM/DIA method at this stage is peak clipping, mainly due to the lack of an API option to define a signal-to-noise ratio for triggering the MSxPRM scan. In some of the measured cases, we observed a delayed triggering of the MSxPRM event, resulting in only partial capture of the endogenous peptide signal.
23	Hall	2023	Our recommendations should be treated with some caution as they might not generalize to situations that are significantly different to our dataset. For example, it is unclear whether our recommendations will apply to these components when elicited by different populations, stimuli, or paradigms, recorded in different environments or by different EEG or electrode systems, or computed by different processing and analytical pipelines.
24	Haven	2023	- This study included many publications by the supervisors, but not all. The number of included first or last author publications for the open science practices varies between 3 and 11; we always included more publications by the supervisor than by the Ph.D. candidate. This meant that at times, we had to exclude pairs because the supervisor did not have a sufficient number of publications, meaning we may have a small bias towards productive supervisors. - We only sampled from four out of eight Dutch UMCs; hence our findings may not generalize to all Dutch UMCs, let alone to other countries. - Our study does not allow for drawing causal inferences on who educated who regarding open science practices.
25	Hayashi	2023	Not explicitly mentioned.
26	Hellen	2023	Not explicitly mentioned.
27	Huckvale	2023	When comparing the performance of the three machine learning models, XGBoost unsurprisingly performed better overall than the Random Forest model while the MLP deep learning method did not improve on the tree-based methods.

			• KEGG is treated as a gold standard, making the assumption that KEGG's description of pathway involvement is complete and without error. This assumption is necessary for training and evaluation and should be somewhat reasonable for central metabolism; however, this assumption is clearly not true, since KEGG is growing. • Since the overall performance as well as the variance in performance are greatly dependent on the pathway category being predicted, certain use-cases may need to exclude certain pathway categories.
28	Jaric	2024	• Similar to multi-center trials in clinical research, multi-laboratory studies might be an ideal solution for preclinical animal studies, but their implementation can be challenging due to logistical demands and intellectual property concerns. • Effective heterogenization of study populations requires the systematic variation of genetic and/or environmental factors that typically vary between independent replicate studies, thereby contributing to between-laboratory variation (that is, heterogeneity in meta-analyses) and thus poor replicability. • The heterogenization of study populations by breeding site did not reduce between-laboratory variability compared to the conventional STA design. Several reasons may explain these unexpected findings. The main reason may be that breeding site contributed only little to total phenotypic variation, much less than test laboratory. • There was little scope for heterogenization by breeding site to reduce between-laboratory variation. Given our previous findings that common environmental differences between animal-rearing facilities can induce persistent phenotypic differences from the molecular to the behavioral level in mice¹⁴, this finding was unexpected. One explanation could be that the rearing conditions in the facilities of professional breeders are much more similar (that is, STA) than the animal facilities of independent research institutions. Furthermore, the six breeding sites belonged to only three breeding companies. Thus, strictly STA operating procedures maintained across different breeding sites within companies could have further reduced phenotypic variation between mice from different breeding sites. Alternatively, the diversity of the mice within breeding sites may have been greater than expected, thereby limiting the scope for variation between breeding sites. This could, for example, be due to variation in age (the age of mice may vary by several days) and origin from different colony rooms. • The pair-housing of male mice could be another factor potentially contributing to larger diversity among mice from the same breeding site. Pair-housing may often result in despotic hierarchies among male mice, and it was found that circulating testosterone levels can differ by up to fivefold between dominant and subordinate males. Such social effects may lead to substantial variability among mice of the same age and strain housed under identical conditions. This may have been further corroborated by the need to single-house some animals for some time before testing due to escalating aggression. Although we accounted for this statistically, it remains possible that the biological effect was more pronounced.
29	Kalvaityte	2024	
30	Kang	2023	• It is difficult to provide general recommendations for design and analysis that apply to all studies, and study design must consider other features in addition to optimizing ESs, such as practical challenges in collecting longitudinal data. • The decision for a particular design or analysis strategy will depend on unknown features of the brain and non-brain measures and their association.
31	Kekecs	2023	• Individual differences cannot be effectively uncovered by this experimental design because of the low number of trials per participant. • The IT audit report noted that the server did not correctly log the authorized login accesses. This was not intended and was only noticed at the end of data collection. This made it impossible to verify that the server access log kept by the PI is accurate and full. • Both research auditors noted that there were some reports of software crashes and Internet connectivity issues reported during the experiments.
32	Kessel	2024	• One threat to validity, therefore, is that the benchmark itself is not optimized for Java and the coding problems might not accurately reflect Java's real world performance.

			• Another threat is the representation and construction of code prompts as part of the coding problems that were used to generate software components. The models under assessment are known to be sensitive to the structure of code prompts, so the threat arises from the fact that the prompts were not optimized for Java code completions (<i>e.g.</i>, Python doctest12 is not supported in Java). • A major challenge in this approach is effectively serializing the observed responses from the software components at run-time (<i>i.e.</i>, serializing Java objects), since information loss may occur as part of the process. • Since EvoSuite often generates tests at the extremes or borders of the input space (<i>e.g.</i>, null values, negative values etc.), there is a chance that some of the recorded component responses are syntactically different, but semantically equivalent. • Due to their tabular form, SRMs obviously cannot be used to store such large, monolithic records.
33	Kolokathis	2024	• A limitation in the current version of EasyMODA is the absence of an automated way to create the workflow image. Another limitation currently is the absence of guidelines on how to complete the MODA for omics based models (proteomics, transcriptomics, Adverse-Outcome Pathway etc).
34	Krahmer	2023	• Our estimates might represent a somewhat lower bound of the true nudging effect, though, as some participants could have been impervious to nudging per se. • In some cases the provision of code turned out to be impeded by the simple fact that such code was never written in the first place. This may be due either to the use of graphical user interfaces (GUI) in statistical analysis tools or to a lack of knowledge about data and code management. Unlike ignorance towards best practices of data and code management, the use of GUI solutions is not inherently problematic as long as sufficient additional information is provided to maintain analytic transparency. • Some of our requests might not have been actively ignored but rather have fallen victim to spam-filters. • We cannot rule out treatment heterogeneity in our experimental results (<i>e.g.</i>, more experienced researchers being more strongly affected by the effort treatment). Authors who received multiple requests from our project (<i>e.g.</i> due to forwarding from other researchers) pose another potential threat to the validity of our conclusions, as these individuals might have been affected by reactivity.
35	Landau	2024	• The representativeness of individuals included in these samples is another critical factor that influences generalizability. Limitations include data storage requirements and data tracking challenges related to curating and tracking large complex image datasets. • MRI-based pipeline is the primary B-PIP approach because of its use of individual anatomy to sample PET tracer uptake. The MRI-based pipeline relies on a co-registered MRI to define ROIs in native space, and for each scan acquired longitudinally, leading to image sampling that is sensitive to morphological characteristics that differ from individual to individual, and from scan to scan over time within each individual. Tau standardization is still an ongoing area of research that involves unique challenges that complicate direct application of the CL approach used for Aβ PET.
36	Liu	2024	• One key challenge is the potential inflation of effect sizes due to the reduced precision inherent in small N studies. This limitation can lead to overestimations of effect magnitude, which may not be replicable in studies with larger and more diverse sample groups. • The studies subjected to replication and our surveyed respondents were not representative samples. Thus, further attempts to assess the generality of our findings are needed. • We do not, by necessity, know the causal impact of policies that encourage large N or effect sizes. Our results do not suggest that large sample sizes are unimportant. While larger sample sizes do improve the precision and generalizability of the effects observed, they also tend to dilute the magnitude of effect sizes, reflecting more modest but potentially more widely applicable effects. Large sample size increases statistical power, and they should be encouraged and incentivized. Instead, our data call for a more comprehensive consideration of issues beyond large N.
37	Li	2024	Not explicitly mentioned.

38	Llovera	2024	- A limitation of this approach of highly standardizing animal models for multicenter studies might be that it underestimates the variability of clinical trials and thereby identifies low efficacy therapeutic candidates that might not be translatable to a clinical stroke population. - Because the Neuroscore and other behavioral tests were not systematically assessed before and after training, we cannot definitively conclude that the intervention influenced the low coefficient of variation among participants posttraining.
39	Luna	2024	- Deep sequencing can be applied to most approaches but is in practice restricted to targeted sequences. Hence, deep sequencing is currently limited to applications that rely on targeted sequencing. In our study, we sequenced all targets molecules in the sample multiple times to maximize the sensitivity. The sensitivity is directly linked to the error rate since a mutation cannot be called with a lower allele frequency than the estimated error rate. - The RT linearity may potentially be improved by optimization of reaction conditions and the use of carriers. Interestingly, we observed differences in capacity to correct errors occurring after RT, i.e., variations in the error correction factor between the various RT conditions. We speculate that this variation may be due to properties of specific reverse transcriptases and their ability to reverse transcribe different RNA sequences as well as RNA modifications. - A low error rate of RT is critical in our approach since the use of UMIs can only correct for errors that occur after cDNA synthesis.
40	Lusa	2024	Not explicitly mentioned.
41	Mager	2023	Not explicitly mentioned.
42	Martinc	2024	In this research, we have focused on the integration and evaluation of open-source taggers from popular NLP libraries, such as NLTK, mostly for reasons of convenience, since these tools have been developed with a specific focus on the ease of integration in larger NLP systems or workflows and were therefore easy to adapt for our specific use case, but also because of the limitations of the TextFlows platform mentioned below.
43	Martin	2023	- One concern is that allowing arbitrary summary queries on a dataset might enable an attacker to reconstruct the data. - Another challenge is handling slowdowns or crashes during resource-intensive analyses due to high compute usage. - Additional challenges include data distribution, network bandwidth, and communication speed.
44	Miyazaki	2023	Not explicitly mentioned.
45	Nance	2024	- First, we focused on plasmode simulations and did not cover alternative approaches, which might be needed if the real data are not available or only partially available. - Second, we emphasized the need to emulate the real data closely but did not discuss how to assess the quality of the emulation. - Third, our presentation did not cover practical implementation, such as how to vary simulation parameters, determine the number of iterations, and parallelize. - Fourth, while our worked examples incorporated common challenges, including confounding, dependence, and missing data, they did not cover other concerns such as generalizability, transportability, and partial identification in detail.
46	Omondi	2024	Despite the authors' best efforts in creating a balanced, full factorial experimental design, there will always be random variation in biological experiments. Fixed effects such as experimental group differences are expected to be generalizable if the experiment is replicated. Random effects (such as gel variation) on the other hand are unpredictable across experiments.
47	Payares-Garcia	2023	One limitation of this work relates to the algorithms and methods implemented in the pipeline. Although many alternatives for each step exist, they are used in precise settings and analysis and are rare in neuroimaging research. NeuroNorm adopts forefront techniques to facilitate data integration and further analysis. NeuroNorm pipeline is also restricted to the Neuroimaging Informatics Technology Initiative (NIFTI) since it is the preeminent format in neuroimaging analysis studies. Likewise, our workflow does not support MRI data from rodents and nonhuman primates. Neuronorm only preprocesses MRI scans from humans. Future work will focus on integrating other preprocessing techniques and formats. The interpretation of MRI outputs processed by NeuroNorm requires careful consideration due to the inherent limitations of the methods incorporated in its pipeline.

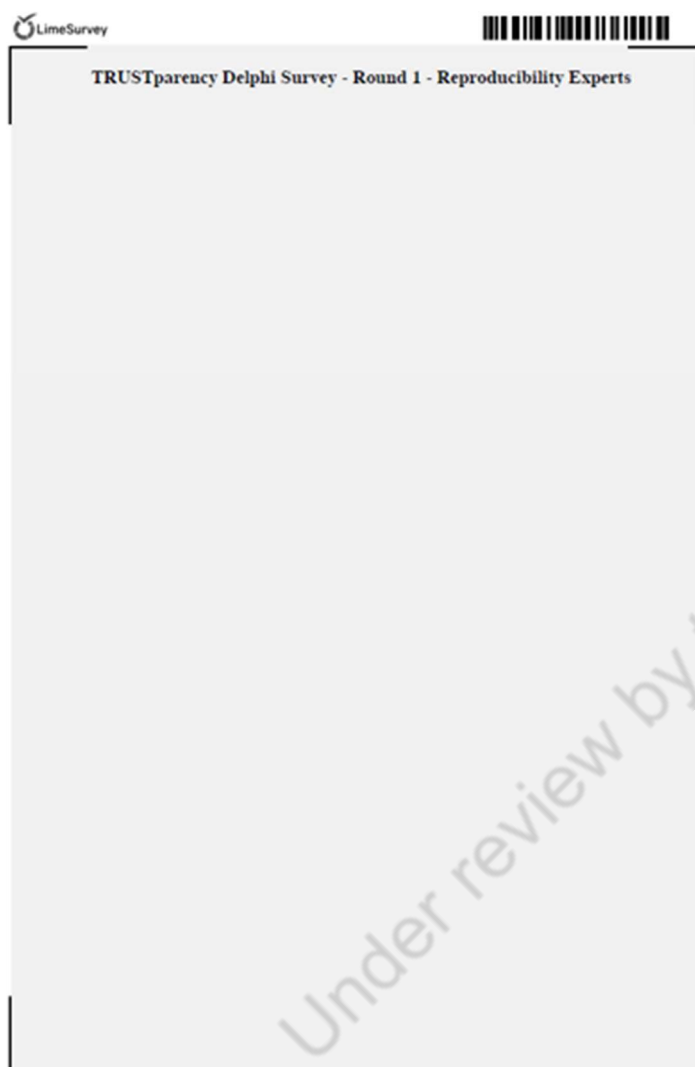
			- Another limitation arises from the HMRF segmentation method deployed in Neuro-Norm. In cases where low-resolution MRI scans are provided, the HMRF algorithm may encounter challenges in identifying appropriate thresholds for classifying gray matter and white matter, potentially resulting in misclassification. - Furthermore, attention must be given to the intensity normalization step in NeuroNorm. The RAVEL method utilized in this process relies on the accuracy of the tissue segmentation results. Furthermore, attention must be given to the intensity normalization step in NeuroNorm. The RAVEL method utilized in this process relies on the accuracy of the tissue segmentation results. Furthermore, attention must be given to the intensity normalization step in NeuroNorm.
48	Pelletier	2024	Not explicitly mentioned.
49	Rehal	2023	Not explicitly mentioned.
50	Rengasamy	2023	- The analysis is that it is not exhaustive, but rather provides an initial step to identifying more reproducible relationships and increasing transparency of analyses, particularly for PIM depression associations. The authors were not able to examine all relevant decisions given the computational complexity and scope of the undertaking for a single, first-of-its-kind manuscript. - As with any other analysis, multiverse analyses are limited by intrinsic study sample characteristics, emphasizing the need of multiverse analyses in different study cohorts to further enhance replicability. - Also, PHQ scores and some covariates were self-reported, though these variables (e.g., past medical illness) have fair concordance with clinical measures.
51	Rieger	2024	One limitation of the method is that the repetition strategy could be computationally demanding on large datasets with a large number of topics. The quality of the results of LDAPrototype in terms of correlation with human perception was not in the scope of the paper, but we plan to investigate it in further studies.
52	Rojas-Velazquez	2024	Not explicitly mentioned.
53	Schacherer	2023	- Using cloud ML services comes with certain trade-offs. Conducting computationally intensive experiments requires setting up a payment account and paying a fee based on the type and duration of the computing resources used. - Furthermore, although the ML services are widely used and likely to be supported for at least the next few years, there is no guarantee that they will be supported in the long term and support the specific configuration of the computing environment used (e.g., software version, libraries). - Independent of the implementation, a major obstacle to the reproducibility of CompPath methods remains their high computational cost.
54	Schuetz	2023	One example is the decision to run the sample recording task locked to the display frame rate (typically 90 Hz) and leave the updating of sensor information to the Vizard runtime. However, hardware such as eye trackers already support faster sampling rates (i.e., 120 Hz for the built-in eye tracker in the Vive Pro Eye HMD), meaning that the eye tracking data may be artificially reduced in fidelity and/or introduce temporal jitter. The same is true for SteamVR controllers and other motion tracking systems. This can be problematic with paradigms that require high temporal accuracy, such as studies employing gaze-contingent stimuli, on-line movement velocity calculations, or highly accurate reaction times.
55	Senior	2023	Not explicitly mentioned.
56	Souza	2023	- First, as we aimed to have comparable software configurations, we did not try to optimize specific parameters for each software. However, specific parameters' optimization could improve their analytical performance. - Also, we did not compare commercial software tools, but we acknowledge their extensive use in the DIA literature. - Finally, our conclusions cannot be generalized to the field outside of lipoprotein and HDL proteomics, since their complexity is different from that of cellular proteome or whole plasma, for example.
57	Sprang	2024	The field of genome-wide data analysis would benefit from studies accurately comparing many possible confounders and their impact on gene expression and reproducibility, although the latter is expected to be low for some diseases.

58	Stahl	2023	- The validated search strategy we used to identify all relevant diagnostic accuracy studies on MEDLINE via PubMed that met our inclusion criteria has a sensitivity of 80.0% and a specificity of 97.3%. It is thus likely that we did not find all eligible diagnostic accuracy studies. - We changed the original STARD checklist by excluding item 11 because a meaningful decision about adherence was not possible. - We only investigated diagnostic accuracy studies published in European Radiology. The last two points may impair the generalizability of our study, but we are still able to draw a conclusion about the quality of reporting of diagnostic accuracy studies in European Radiology for 2015 and 2019. - Although the STARD statement comes along with detailed explanations and is highly elaborated, there remains some subjectivity in scoring the items. Our scoring method could have a negative effect on items that need to be described in more detail such as item 6 (eligibility criteria) or item 10b (reference standard, in sufficient detail to allow replication).
59	Tang	2024	Not explicitly mentioned.
60	Tång-Hallbäck	2025	Not explicitly mentioned.
61	van Houdt	2023	- A limiting factor was the availability of code in the Python language. - Ground-truth reference values are often difficult to define. (...) Thus, the testing framework, while designed to assure an acceptable level of quality and to compare the quantitative outputs of different code contributions, is not intended as a means to rank or recommend specific contributions. - Adding new code contributions to the current testing framework requires some manual intervention by taskforce members or the user themselves. - Speed and robustness (for example, to invalid inputs) were not assessed at this stage. Nor did we review the code in a line-by-line manner to detect errors (although we note that our testing framework did detect a small number of coding errors that were corrected with permission from the contributing authors). - The current test data does not cover parameter ranges relevant to all applications. - We focused on test cases where there is a reasonable expectation of obtaining a valid result.
62	Van Kampen	2024	- ENCORE is a compromise based on previous ways of working and, therefore, may not always fit the preferred way of working of an individual researcher. - The current sFSS Navigator has limited functionality, ways to present information and configuration options. - A shortcoming of ENCORE is the lack of a clear and easy approach to specify explicit links between results, code, data, and concepts. - Another challenge that is only partially addressed by ENCORE concerns the preservation of the full computing environment. - ENCORE projects can easily be shared but currently we do not have a good mechanism to remove parts that should not be shared such as sensitive (patient) data, controlled access data obtained from public repositories, or PDF copies of copyright protected scientific publications. - ENCORE requires that all data used by the computations are within the sFSS structure.
63	Viant	2024	Not explicitly mentioned.
64	Wallace	2024	Not explicitly mentioned.
65	Wirz	2024	Not explicitly mentioned.
66	Yousefirizi	2024	Not explicitly mentioned.
67	Zhang	2023	Not explicitly mentioned.
68	Zhu	2023	- The reproducibility of a GSE is closely tied to the evaluation criteria. There are still many questions about how to evaluate an experiment is reproducible. - From our perspective, there are still some shortcomings with regard to the ODD protocol, TRACE notebook, and Minimum Information Required in the Annotation of Biochemical Models (MIRIAM), our study aims to learn from the experience of existing guidelines, as well as to reference and welcome geo-communities to provide feedback by contacting the lead author.

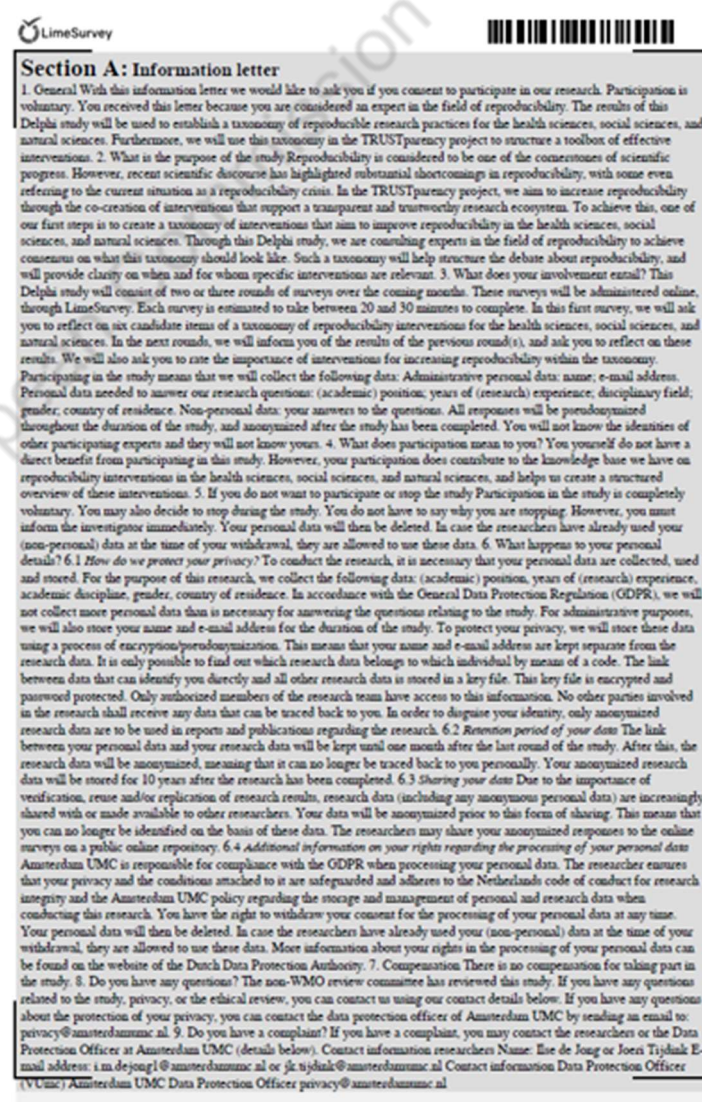
			• One important concern in implementing a documentation strategy is a lack of tools to facilitate the recording processes.
--	--	--	--

Under review by the European Commission

5.8 Delphi round 1 survey



Page 1



Page 2

LimeSurvey

Section B: Informed consent

Condition: I have read and understood the information about the project, as provided in the participation information document; I have been given the opportunity to ask questions about the project and my participation and I have been given sufficient time to decide whether I want to participate; I voluntarily agree to participate in the research project and I understand I can end my participation at any time without giving a reason for it; I understand that the researchers will collect and use survey data in the way that is described in the information document; I understand that until one month after the final round of data collection, my personal data, which are obtained for administrative purposes only, will be kept in pseudonymized form; I understand that my research data within this research will be obtained for scientific purposes and will be available in anonymized form for verification, reuse and replication in a repository for 10 years; I understand that the researchers may present some of the findings publicly at conferences or in publications.

B1.

I consent to participate in this study

Yes ☐

No ☐

Section C: Introduction

Process of the Delphi In the next pages, we will ask your opinion on several questions. These questions can be answered on a 5-point scale ranging from either "1. Strongly disagree" to "5. Strongly agree", or from "1. Not relevant at all" to "5. Extremely relevant". If you feel like you do not have the right expertise to answer a question, you can use the "No expertise" option. After each question we ask you to formulate the rationale behind your answer in an open text field. All open text fields are optional. However, your arguments, ideas, suggestions and feedback are highly valuable to further improve the taxonomy and address our study aims. Please do not include any personally identifying information about people or institutions. All personally identifying information will be deleted from our data. Consensus We will consider to have reached consensus for a question when 67% or more of the panelists have scored it 4 or 5. Based on the argumentation and rate of consensus among the panelists, the steering committee, consisting of the researchers from Amsterdam UMC, will modify, include new, or exclude elements and/or their descriptions. This will be presented in the next round and the steering committee will communicate all decisions made. Notes on terminology Before we start, we first provide you with some notes on terminology, as the concept of reproducibility remains ambiguously and inconsistently defined across research fields. When we talk about reproducible research practices, we refer to practices that enhance the reproducibility of research. When we talk about reproducibility, we refer to the extent to which the results of a replication agree with those of the original study. Lastly, when we talk about replication, we refer to an actual attempt to re-produce an earlier finding or experiment. Let's get started! On the following pages, you will see descriptions of stages of the research cycle. For each stage, we ask you to tell us the extent to which you agree with the given description, and rate the relevance of this stage for facilitating reproducible research practices. We also ask you to formulate the rationale for your answers, and suggest any changes and/or additions to the descriptions. You will be asked about the following stages: Formulate a hypothesis/research question; Design of the study; Data collection; Data analysis; Report the findings; and Post-publication. At the end of the survey, we will ask a few questions on demographics.

LimeSurvey

Section D: Stage: Formulate a hypothesis/research question

Description: The early stage in which ideas are formed, the contours of the study are described and resources are secured. Includes: Specify research questions/hypotheses; Develop theoretically or empirically grounded research questions/hypotheses. Sketch the contours of the study plan. Obtain resources: Preparing applications to secure financial, human, and technical resources.

D1. How strongly do you agree that this description appropriately captures the activities involved in the "Formulate a research question/hypothesis" stage of the research cycle?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

D2. How relevant is the "Formulate a research question/hypothesis" stage of the research cycle for facilitating reproducible research practices?

1. Not relevant at all ☐

2. Slightly relevant ☐

3. Moderately relevant ☐

4. Very relevant ☐

5. Extremely relevant ☐

No expertise ☐

D3. If you would like to share any changes you would make to the description of the "Formulate a research question/hypothesis" stage of the research cycle, please do so here:

LimeSurvey

D4. If you would like to share any additions you would make to the topics included in the description of the "Formulate a research question/hypothesis" stage of the research cycle, please do so here:

D5. If you would like to share the arguments for your rating of relevance for the "Formulate a research question/hypothesis" stage of the research cycle, please do so here:

Section E: Stage: Design of the study
 Description: The planning stage in which methodological decisions are made, operationalized, and documented. Includes: Develop study protocol: Creating a detailed plan for how the study will be conducted, including sampling, measurement instruments, data analysis plan and data management plan. Operationalize the study protocol: Obtaining or creating the materials, instruments, tools, or software needed to execute the study. Uploading the study protocol in a time-stamped repository before the start of data collection.

E1. How strongly do you agree that this description appropriately captures the activities involved in the "Design of the study" stage of the research cycle?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

LimeSurvey

E2. How relevant is the "Design of the study" stage of the research cycle for facilitating reproducible research practices?

1. Not relevant at all ☐

2. Slightly relevant ☐

3. Moderately relevant ☐

4. Very relevant ☐

5. Extremely relevant ☐

No expertise ☐

E3. If you would like to share any changes you would make to the description of the "Design of the study" stage of the research cycle, please do so here:

E4. If you would like to share any additions you would make to the topics included in the description of the "Design of the study" stage of the research cycle, please do so here:

E5. If you would like to share the arguments for your rating of relevance for the "Design of the study" stage of the research cycle, please do so here:

LimeSurvey

Section F: Stage: Data collection
 Description: The execution stage of the research, involving data collection, storage, and documentation. Includes: Collect data: Executing the data-gathering process (e.g., running experiments, conducting interviews, administering surveys). Store data: Document procedures: Keeping detailed records of how the study was conducted, including any deviations from the study protocol and decisions made during execution.

F1. How strongly do you agree that this description appropriately captures the activities involved in the "Data collection" stage of the research cycle?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

F2. How relevant is the "Data collection" stage of the research cycle for facilitating reproducible research practices?

1. Not relevant at all ☐
 2. Slightly relevant ☐
 3. Moderately relevant ☐
 4. Very relevant ☐
 5. Extremely relevant ☐
 No expertise ☐

F3. If you would like to share any changes you would make to the description of the "Data collection" stage of the research cycle, please do so here:

LimeSurvey

F4. If you would like to share any additions you would make to the topics included in the description of the "Data collection" stage of the research cycle, please do so here:

F5. If you would like to share the arguments for your rating of relevance for the "Data collection" stage of the research cycle, please do so here:

Section G: Stage: Data analysis
 Description: The stage where collected data is prepared, processed, and analysed to answer the research question. Includes: Prepare and clean data: Organize, transform, and clean raw data to prepare it for analysis. Run analyses: Conduct statistical or qualitative analyses. Validate results: Assess the robustness and reliability of findings.

G1. How strongly do you agree that this description appropriately captures the activities involved in the "Data analysis" stage of the research cycle?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

LimeSurvey

G2. How relevant is the "Data analysis" stage of the research cycle for facilitating reproducible research practices?

1. Not relevant at all ☐

2. Slightly relevant ☐

3. Moderately relevant ☐

4. Very relevant ☐

5. Extremely relevant ☐

No expertise ☐

G3. If you would like to share any changes you would make to the description of the "Data analysis" stage of the research cycle, please do so here:

G4. If you would like to share any additions you would make to the topics included in the description of the "Data analysis" stage of the research cycle, please do so here:

G5. If you would like to share the arguments for your rating of relevance for the "Data analysis" stage of the research cycle, please do so here:

LimeSurvey

Section H: Stage: Report the findings

Description: The stage where findings are interpreted and shared with the (scholarly) community. Includes: Interpret findings: Contextualize the results in relation to the research question, hypothesis, theory, and previous literature. Write report: Report the findings of the study. Publish/disseminate findings: Submit the manuscript to a preprint server, journal or conference. Share data/code/materials: Make datasets, analysis scripts, and research materials available for accountability or reuse purposes.

H1. How strongly do you agree that this description appropriately captures the activities involved in the "Report the findings" stage of the research cycle?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

H2. How relevant is the "Report the findings" stage of the research cycle for facilitating reproducible research practices?

1. Not relevant at all ☐

2. Slightly relevant ☐

3. Moderately relevant ☐

4. Very relevant ☐

5. Extremely relevant ☐

No expertise ☐

H3. If you would like to share any changes you would make to the description of the "Report the findings" stage of the research cycle, please do so here:

LimeSurvey

H4. If you would like to share any additions you would make to the topics included in the description of the "Report the findings" stage of the research cycle, please do so here:

H5. If you would like to share the arguments for your rating of relevance for the "Report the findings" stage of the research cycle, please do so here:

Section I: Stage: Post-publication
 Description: The post-publication stage in which findings are assessed, scrutinized, and extended. Includes: Peer-review and commentary: Participating in or responding to peer review processes that critically assess the study's design, analysis, and interpretation. Corrections: Addressing errors, omissions, or limitations discovered after publication by issuing corrections, retractions, or clarifications. Replication: Conducting independent attempts to reproduce the findings of the study using the same or closely related methods and materials.

I1. How strongly do you agree that this description appropriately captures the activities involved in the "Post-publication" stage of the research cycle?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

LimeSurvey

I2. How relevant is the "Post-publication" stage of the research cycle for facilitating reproducible research practices?

1. Not relevant at all ☐

2. Slightly relevant ☐

3. Moderately relevant ☐

4. Very relevant ☐

5. Extremely relevant ☐

No expertise ☐

I3. If you would like to share any changes you would make to the description of the "Post-publication" stage of the research cycle, please do so here:

I4. If you would like to share any additions you would make to the topics included in the description of the "Post-publication" stage of the research cycle, please do so here:

I5. If you would like to share the arguments for your rating of relevance for the "Post-publication" stage of the research cycle, please do so here:

LimeSurvey

Section J: That concludes the stages of the research cycle! If there is anything else you would like to share, please do so here:
On the next page, we will ask you to answer some final questions on demographics.

J1.

Section K: Demographics

K1. Which discipline do you primarily work in?
Select the one that best describes your main area of work

Health sciences ☐
Social sciences ☐
Natural sciences ☐
Other ☐

K2. Please specify:

K3. What is your gender?

Female ☐
Male ☐
Non-binary/third gender ☐
None of the above ☐
Prefer not to disclose ☐

K4. In which country do you primarily work?
Please enter the country where you are primarily based for work

LimeSurvey

K5. How would you rate your experience with research reproducibility (researching or implementing reproducibility practices)?

Not at all experienced ☐
Slightly experienced ☐
Moderately experienced ☐
Very experienced ☐
Extremely experienced ☐

K6. What best describes your current role?

Researcher ☐
Research support staff (e.g., data steward, librarian, IT support) ☐
Research manager or administrator ☐
Journal editor or publishing professional ☐
Member of reproducibility network ☐
Member of a learned society or professional association ☐
Policy maker (e.g., at a funding agency or government body) ☐
Other ☐

K7. Please specify "Other":

Section L: This is the end of round 1!

L1.

In a few weeks, we will contact you again for round 2 of this Delphi study. In case you do NOT want to be contacted again, please tick the box below. If you are happy to be contacted again, you may skip this question.

I do NOT want to be contacted about this study again ☐

Thank you for your participation! This is the end of the survey. Unless you indicated that you do not want to be contacted for future rounds of this Delphi study, we will contact you within the next few weeks about the next round of the survey.

5.9 Feedback report for round 1

A Delphi Consensus Study to Taxonomize Reproducibility Interventions: Round 1 Feedback Report

Ilse M. de Jong, Joeri K. Tijdkink, Mariska M. G. Leeflang, Lex Bouter,

Panagiotis Kavouras, Mads P. Sørensen

5.9.1 Overview of round 1

This paragraph provides a brief overview of the aims of our Delphi study and its procedure. The preregistration for this study can be found [here](#). The Delphi method consists of a series of sequential questionnaires or “rounds” where a panel of experts is asked to provide their judgements on a topic (Linstone & Turoff, 1975).

The aim of our Delphi study is to obtain consensus among reproducibility experts on a taxonomy of interventions aimed at improving reproducibility (i.e., to make it easier to replicate previous studies or to reproduce the results of previous studies) related to different phases of the research cycle. A taxonomy is a structured classification system that helps organize complex information into categories based on shared characteristics. In this case, we aim to create a taxonomy of reproducibility interventions using the research cycle as a framework. Such a taxonomy will make it easier for different stakeholders to identify relevant interventions at each stage of the research cycle.

The relevance of the different phases of the research cycle, and of the research cycle itself, can vary across disciplines. For this reason, we focus specifically on the natural sciences, social sciences, and health sciences in this Delphi study.

In the first round, participants were presented with descriptions of six different stages of the research cycle. They were asked to rate the extent to which they agreed with each description on a 5-point scale, and to provide comments, suggest improvements, and propose alternative stages, along with justifications for these suggestions. For each stage, participants were also asked to rate the relevance for facilitating reproducible research practices on a 5-point scale and they were given the option to provide arguments for their rating.

5.9.2 Demographics

Response rate

We invited 228 panelists to participate in the first round of the survey. These panelists were identified by our consortium members to have expert knowledge on reproducibility. These invited experts mostly included people involved in other reproducibility projects and members of national reproducibility experts. Of the invited panelists, 72 finished the survey, leading to a response rate of 32%. Twenty-three additional participants started, but did not complete the survey. Of these 23 participants, 18 ended on or before page 4 of the survey, which held the explanation of the Delphi survey, meaning that they did not answer any of the survey questions besides informed consent. Three non-completers only answered the questions relating to the first stage of the research cycle, one also answered questions relating to the second stage, and one stopped after completing the questions relating to the third stage. As none of the panelists who partially completed the survey

filled in more than half of the survey, we decided to not consider their responses in analyzing the data.

We also distributed the survey through the LinkedIn pages of the TRUSTparenCy and TIER2 projects. Two of the invited panelists also shared the survey in their network. These non-personal invites led to five extra participants, who completed the survey. Of these five participants, two left their e-mail addresses so they can be contacted for future rounds. Sixteen people opened the survey, but did not complete it. One of these people answered the survey questions up until the second stage of the research cycle. All others stopped the survey earlier. We decided not to include the answers of partial responders in data analysis. Table 1 shows an overview of the response rate for both the personal invites and the open invites.

Table 1: Overview of response rate

Survey	Number of invites	Started round 1	Partially completed	Finished round 1	Response rate
Personal invite	228	95	23	72	32%
Open invite	-	21	16	5	-

Demographics panelists

Table 2 presents the demographics of the panelists who finished round 1 of the Delphi.

Table 2: Demographics of the panelists who finished round 1

Category	Nr	Percentage
Gender		
Female	35	45%
Male	41	53%
Non-binary	0	0%
Prefer not to disclose	1	1%
None of the above	0	0%
Geographic region		
Africa	3	4%
Asia	0	0%
Europe	62	81%
Latin America	5	6%
North America	5	6%
Oceania	2	3%
Discipline		
Health Sciences	37	48%
Natural Sciences	12	16%
Social Sciences	22	29%
Other	6	8%
Experience with reproducibility		
Not experienced at all	1	1%
Slightly experienced	4	5%
Moderately experienced	29	38%
Very experienced	30	39%
Extremely experienced	13	17%

Role*		
Researcher	62	81%
Research support staff	8	10%
Research manager or administrator	10	13%
Journal editor or publishing professional	10	13%
Member of reproducibility network	41	53%
Member of learned society	3	4%
Policy maker	3	4%
Other	1	1%

* Participants could select more than one role.

5.9.3 Consensus rates after round 1

For each of the six presented stages of the research cycle, panelists were asked to answer the following two questions: “To what extent do you agree...” and “How relevant...” on a 5-point Likert scale ranging from 1. Strongly disagree/ Not relevant at all to 5. Strongly agree/ Extremely relevant. For all items consensus was reached based on our predefined rates of consensus of 67% of panelists strongly agreeing or agreeing to a question, or rating the relevance as very or extremely relevant. Rates of consensus ranged from 71% to 97%. Table 3 shows all ratings of consensus.

Table 3: Rates of consensus and number of comments for all phases of the research cycle

Phase	Rate of consensus		Number of comments		
	Description	Relevance	Changes to description	Addition to description	Arguments for relevance
Formulate a RQ	71%	71%	32	33	34
Design study	82%	95%	29	27	24
Collect data	86%	94%	24	24	20
Analyze data	88%	97%	25	21	17
Write report	82%	92%	22	20	24
Post-publication	79%	71%	28	18	19

Note. This table shows the rate of consensus for the description and relevance of each phase of the research cycle, as well as the number of comments given in response to the open textboxes after round 1 (N = 77). “Consensus” is defined as the percentage of participants who answered either “strongly agree” or “agree” (for the description) or who answered either “extremely relevant” or “very relevant” (for the relevance). RQ = research question.

5.9.4 Comments

On average, we received 24 comments for each open-ended question that we asked. These included valuable and insightful comments on the given descriptions of the stages of the research cycle, and arguments for each stage’s relevance. All comments were analyzed pseudonymously and grouped on type of answer and overlapping content between answers. Data analysis was done manually by IJ, and checked by JT and ML. Summaries of the responses given to each question and selected example comments will be provided in the next sections.

5.9.5 The steering committee

The steering committee consisted of researchers from the Amsterdam UMC team (IJ, JK, ML, LB). All decisions were checked by the other members of the research team (PK, MS).

5.9.6 Stage 1: Formulate a research question/ hypothesis

Description

Proposed description

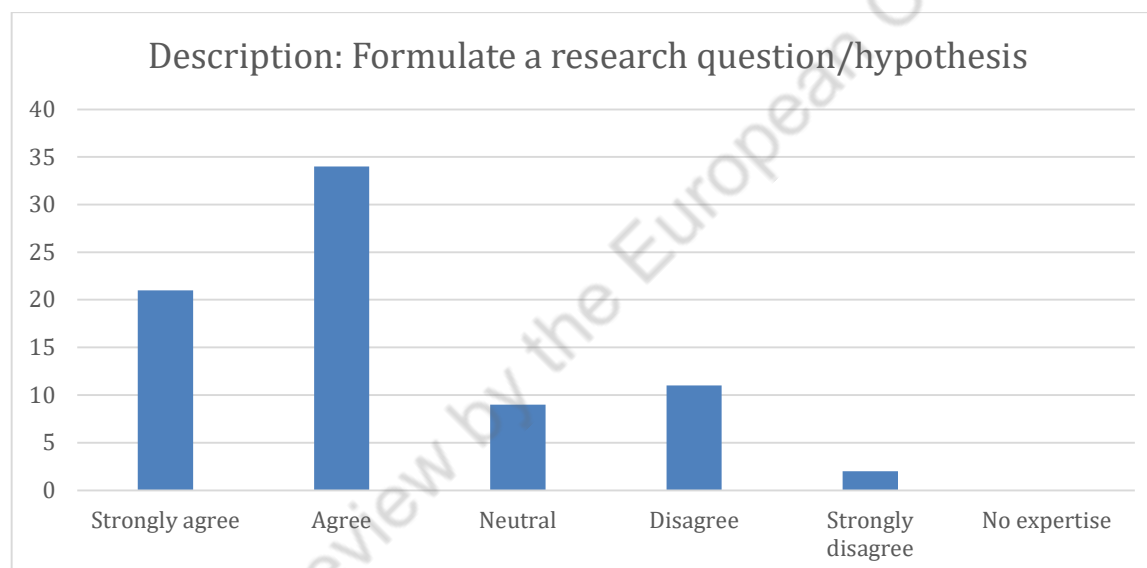
The early stage in which ideas are formed, the contours of the study are described and resources are secured.

Includes:

- Specify research questions/hypotheses: Develop theoretically or empirically grounded research questions/hypotheses.
- Sketch the contours of the study plan.
- Obtain resources: Preparing applications to secure financial, human, and technical resources.

Question: How strongly do you agree that this description appropriately captures the activities involved in the "Formulate a research question/hypothesis" stage of the research cycle?

Consensus: 71%



Summary of comments

Prompt: If you would like to share any changes you would make to the description of the "Formulate a research question/hypothesis" stage of the research cycle, please do so here:

Comments: 32

Summary: In 15 comments, panelists mentioned that they were unsure whether obtaining resources is part of this stage. Five panelists specifically mentioned they see this as a separate or later phase (see comment 1.1 below). Six panelists did not specify why they believe obtaining resources is not part of this stage. Three panelists suggest that the issue may be with the header of this stage ("Formulate a research question/hypothesis"), and suggest to broaden the phase's header (see comment 1.2). One panelist said they believe obtaining resources is an administrative task, rather than a scientific task (see comment 1.3).

Eight participants commented on the inclusion of "sketch the contours of the study plan". Five of these said that this is not part of this stage, and could perhaps be moved to a different stage (see comment 1.4). Three panelists specifically mentioned that they find the word "contours" too vague or weak.

While additions to the description were explicitly requested in the next prompt, six panelists already mentioned here that they believed doing a literature review or identifying the gap in the literature is missing.

Prompt: If you would like to share any additions you would make to the topics included in the description of the "Formulate a research question/hypothesis" stage of the research cycle, please do so here:

Comments: 33

Summary: Eighteen participants wrote that a necessary part of this stage of research is doing a literature review and identifying the gap in the literature (see comment 1.5). Five participants suggested to add an assessment of the feasibility of the study (see comment 1.6). Two participants suggested to explicitly mention getting ethical approval.

Selection of comments

Comment 1.1: I don't think I'd include "securing resources" within the "formulating a research question" stage. To me, these are separate issues. You can formulate a hypothesis without securing resources or secure resources for hypotheses that have been formulated long ago by others. We may want to have securing resources as a part of the research process, but I don't think it should be in the same stage of formulating a hypothesis.

Comment 1.2: I am not sure if securing resources fits well with the title of "formulate a research question/hypothesis", maybe the title need to be adjusted to be a bit more broad as research proposal for example.

Comment 1.3: In my opinion securing resources is not part of this stage, as it is very much an administrative task and not a scientific task. I also think that sketching the research plan comes later.

Comment 1.4: I would take forming the "contours of the study" to be part of the design phase than part of coming up with the research question. But I am not sure what exactly is meant with contours of the study. If just the basic idea, then maybe somewhat relevant.

Comment 1.5: Desk research of existing studies on the topic to anchor theoretically and to make sure there is added value of the research.

Comment 1.6: I would add a point about estimating, even roughly, the feasibility of the project, based on local knowledge or example in the literature.

Response of steering committee

Many panelists commented on the inclusion of securing resources in this stage, and suggest that it could be moved to its own separate stage or to the next stage (design of the study). Part of the reason that we included this stage, is that there are interventions to foster reproducible research practices that are aimed at the earliest stage of the research cycle. These include interventions implemented

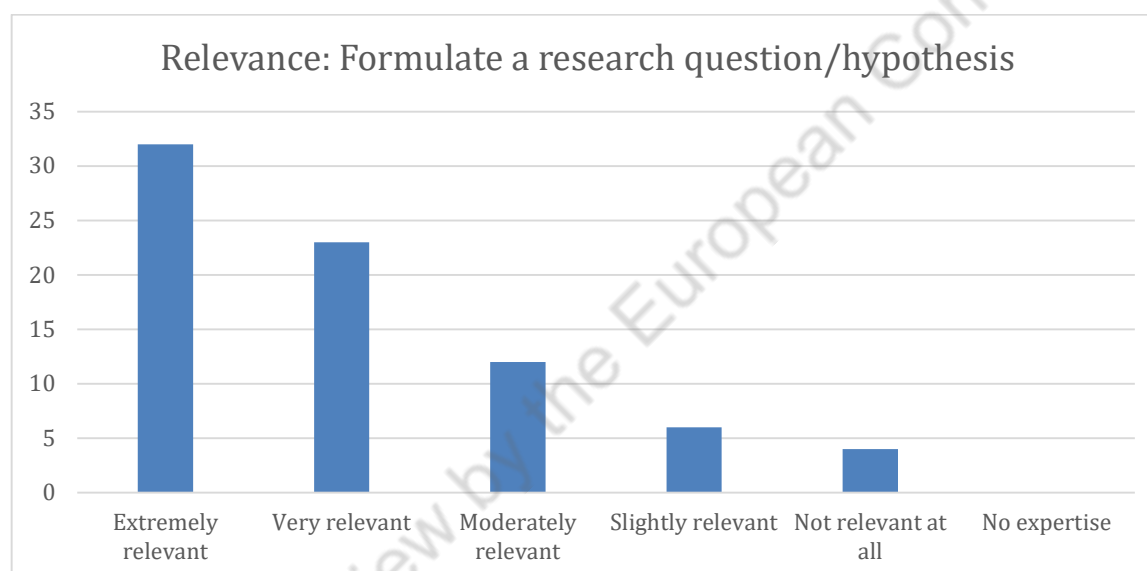
by funders, such as a mandate for writing a data management plan, or making data openly available. While these may be executed by researchers at later stages, they are implemented before the research cycle really starts. However, based on the comments that panelists made, we have realized that these interventions have a different

status than interventions that are implemented by researchers themselves, at different stages of the research cycle. Interventions such as those implemented by funders lay the foundation for further research. They therefore perhaps do not belong within the research cycle, but in a foundational, pre-research cycle stage. The same seems to go for certain types of training or infrastructure. These do not necessarily belong in one stage of the research cycle, but rather build the structure that is needed to facilitate the engagement in reproducible research practices throughout the research cycle.

Relevance

Question: How relevant is the "Formulate a research question/hypothesis" stage of the research cycle for facilitating reproducible research practices?

Consensus: 71%



Summary of comments

Prompt: If you would like to share the arguments for your rating of relevance for the "Formulate a research question/hypothesis" stage of the research cycle, please do so here:

Comments: 33

Summary: Out of 33 comments, 16 gave an argument for why this stage is relevant for facilitating reproducible research practices. Most of these argue that this stage lays the foundation for future steps of the research (see comment 1.7). Six comments included arguments for why this stage may not be (as) relevant for facilitating reproducible research practices, mostly relating to how being able to reproduce someone's thought process in formulating a research question may not be the most relevant for reproducibility (see comment 1.8). Seven comments included nuanced views on the relevance of this stage, arguing that it depends on aspects like the scope or goal of the study (see comments 1.9 and 1.10). Four panelists referred to their answer(s) given to the previous question(s), but these answers did not give a clear indication on high or low relevance.

Selection of comments

Comment 1.7: I think that actions aimed at the very early stages of a research project have particular importance for improving reproducibility and trustworthiness. If we can strengthen good research practice right at the start of the project, it could cascade and strengthen the later parts of the project too.

Comment 1.8: Again, I see hypothesis formulation as a distinct section that precedes study planning. To my mind, it is much more based on literature reviews and “brainstorming” or idea formulation, from where we proceed to study planning. As such, it might be difficult (or night impossible) to actually copy someone’s thought process.

Comment 1.9: The answer to the question depends on whether we are considering the reproducibility of a single research study, or a line of research. For a single research study, information about how the question was developed is not essential for reproducibility. For a collection of research on a particular topic, information on how the research question was developed may be more important for reproducibility.

Comment 1.10: The research question connects to reproducibility only if there is a goal to reproduce something. If it is discovery-based research, I think reproducibility might even be an obstacle.

Response of steering committee

The question on the relevance of the “Formulate a research question/hypothesis” stage of the research cycle for facilitating reproducible research practices leads to nuanced answers. On the one hand, this step may lay the foundation for future steps, but at the same time, being able to reproduce someone’s thought process in formulating a research question may not be as relevant for reproducibility. For replication studies that focus on the same research question, whether the research question is relevant, makes any sense and how it arose, are arguably even completely irrelevant. These factors might inform the decision whether the initial study is a priority case for replication, but that’s an issue that is irrelevant for creating the taxonomy.

5.9.7 Stage 2: Design of the study

Description

Proposed description

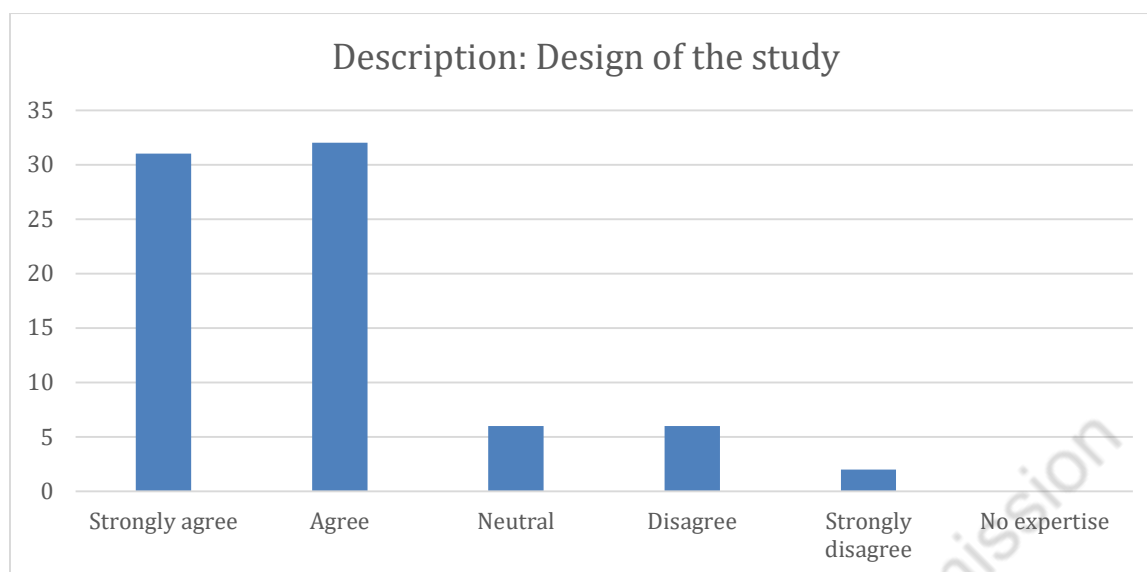
The planning stage in which methodological decisions are made, operationalized, and documented.

Includes:

- Develop study protocol: Creating a detailed plan for how the study will be conducted, including sampling, measurement instruments, data analysis plan, and data management plan.
- Operationalize the study protocol: Obtaining or creating the materials, instruments, tools, or software needed to execute the study.
- Uploading the study protocol in a time-stamped repository before the start of data collection.

Question: How strongly do you agree that this description appropriately captures the activities involved in the “Design of the study” stage of the research cycle?

Consensus: 82%



Summary of comments

Prompt: If you would like to share any changes you would make to the description of the "Design of the study" stage of the research cycle, please do so here:

Comments: 29

Summary: Eleven panelists commented that they believe uploading the study protocol in a time-stamped repository should be optional but recommended, and is not a necessary step of the research cycle (see comment 2.1). Seven panelists noted that the wording could be changed to promote inclusivity of different types of research (see comment 2.2). In three comments, panelists suggest that "obtain resources" from the previous stage might fit better in this stage (see comment 2.3).

Prompt: If you would like to share any additions you would make to the topics included in the description of the "Design of the study" stage of the research cycle, please do so here:

Comments: 27

Summary: There was considerable variation in the comments that panelists gave in response to this prompt. In five comments, panelists suggested to include pilot testing (see comment 2.4). Six panelists emphasized the importance of data management plans, statistical analysis plans, and/or project management plans (see comment 2.5). Five panelists suggested a (minor) change in wording for one of the subcategories to improve the applicability to different disciplines or research methods. In two comments, panelists suggested that updating the plan when changes are made during the research process should also be included here. Four panelists suggested the addition of another item, but these were all mentioned only once. Suggestions made were: assembling and training the research team, documenting the extent to which the research is a (partial) replication of previous work, clearly defining materials, and identifying how research results will be communicated to the public.

Selection of comments

Comment 2.1: Uploading the study protocol in advance is not strictly a requirement for a reproducible research as long as it is documented correctly.

Comment 2.2: The section as written seems to focus on studies that are collecting new data, changes could be made to reflect secondary data analysis projects or evidence synthesis projects.

Comment 2.3: “Obtaining of creating the materials, instruments, tools, or software needed to execute the study” seems to partly overlap with securing resources in phase 1 (where I think it does not fit well). Perhaps move all aspects of securing resources to this phase?

Comment 2.4: Pilot testing – Planning small-scale testing of instruments or procedures to refine methods before full implementation.

Comment 2.5: For statistical analysis plan, defining any assumption checks, sensitivity analyses (i.e., testing the robustness of assumptions), subgroup analyses (are there differences in biologically plausible subgroups of the population that may have an effect on the outcome of interest), how missing data are handled. How is data being validated and ensuring its completeness is also a key component of the quality assurance and data management plan.

Response of steering committee

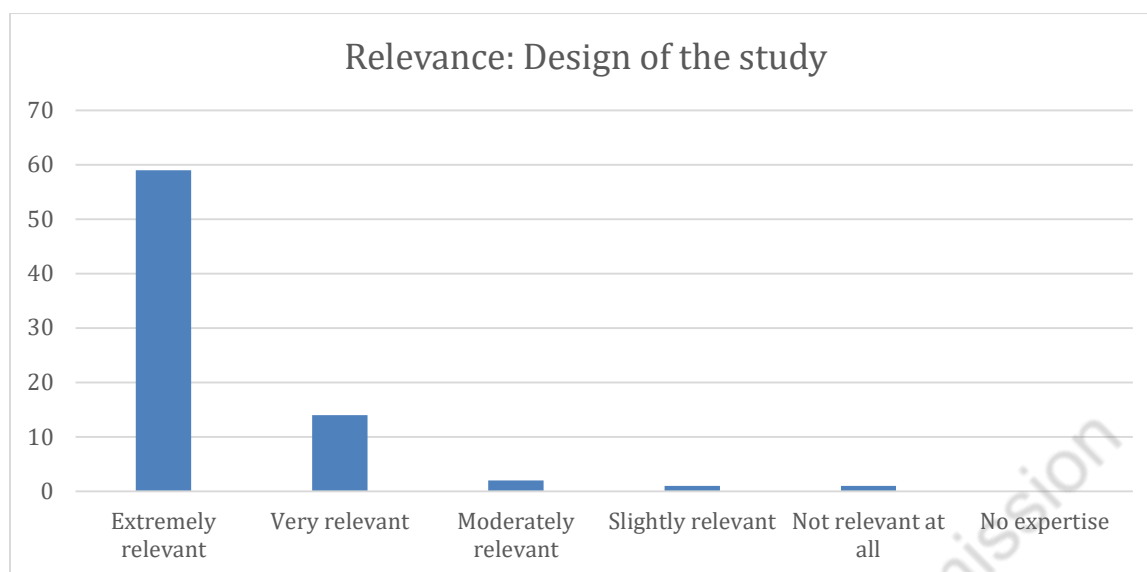
As noted by multiple panelists, there is considerable overlap between this stage of the research cycle and the previous stage. This goes especially for securing resources (originally in stage 1) and obtaining materials (originally in stage 2). As both of these stage relate to the preparational stage of doing research, we may have to see this as one stage, as opposed to two separate stages.

The most frequently made comment in response to this description was that uploading the study protocol in a time-stamped repository should be optional, or may not apply to every type of research. While the most important thing regarding the writing of a study protocol is indeed arguably not the actual uploading in a time-stamped repository, this is still an important and relevant step for preregistration. We want to stress that for the taxonomy, it is not necessary that every category of interventions applies to every type of research. At the same time, we do realize that we may need to change some wordings to improve the applicability of the different categories to different ways of doing research, as suggested by multiple panelists.

Relevance

Question: How relevant is the "Design of the study" stage of the research cycle for facilitating reproducible research practices?

Consensus: 95%



Summary of comments

Prompt: If you would like to share the arguments for your rating of relevance for the "Design of the study" stage of the research cycle, please do so here:

Comments: 24

Summary: Thirteen comments include an argument for why this stage is relevant for facilitating reproducible research practices. The most frequently given argument is that this stage is crucial for transparency and reliability (see comment 2.6). No comments argue for why this stage is not relevant for reproducible research practices. However, five panelists give a more nuanced view of the relevance of this stage. Four of these argue that the relevance depends on the discipline and type of research (see comment 2.7). One panelist notes that having information on the design of the study is important, but that reproducing the process of designing the study is not (see comment 2.8). Five panelists left a comment that did not specify an argument.

Selection of comments

Comment 2.6: The relevance of this stage for reproducible science cannot be underestimated. By thinking out and operationalising your study and uploading it to a repository in advance you ensure that transparency is possible. But, this takes extra time (which we do not always have of our institutional obligations and traditional research culture).

Comment 2.7: Uploading the study protocol in a time stamped repository is relevant only for some disciplines (e.g. health) yet not for all.

Comment 2.8: This is nuanced; I think actually that study planning, in the sense of developing a protocol, calculating a sample size, and preregistering a protocol are extremely important. Whether this can actually be reproducible, as defined here, might be questionable, as with the previous question. How can I reproduce this process, for example: "Operationalize the study protocol: Obtaining or creating the materials, instruments, tools, or software needed to execute the study." What would it mean to reproduce someone's process of obtaining (i.e., acquiring, not using) materials? I think the point is to have the information on these materials so that a study can be reproduced, not to reproduce the process of obtaining these materials.

Response of steering committee

Consensus for the relevance of this stage for facilitating reproducible research practices was very high (95%). As said by one of the panelists, we agree that “the relevance of this stage for reproducible science cannot be underestimated”.

5.9.8 Stage 3: Data collection

Description

Proposed description

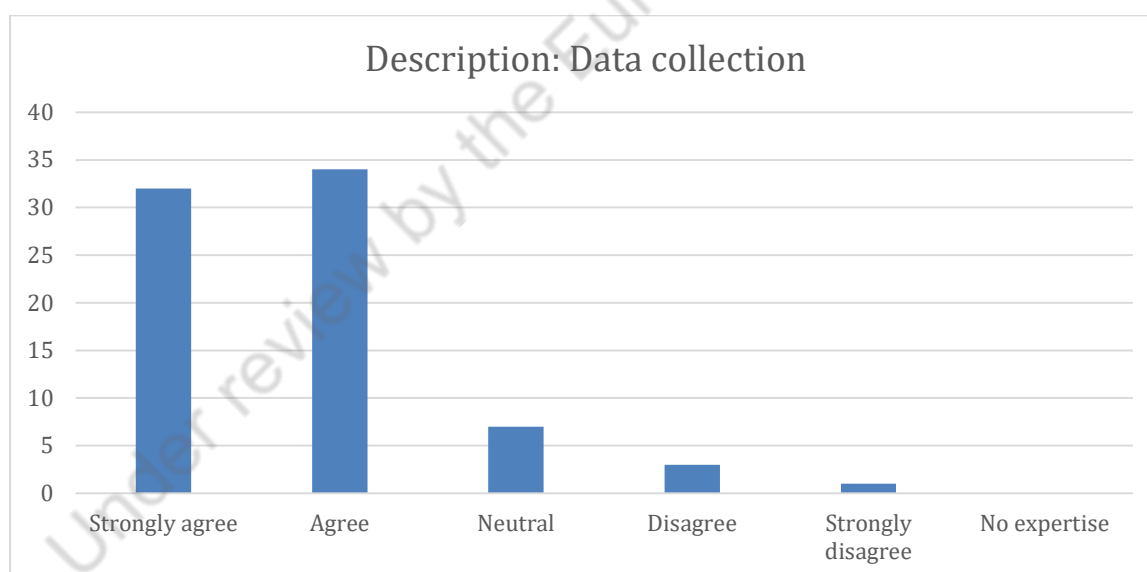
The execution stage of the research, involving data collection, storage, and documentation.

Includes:

- Collect data: Executing the data-gathering process (e.g., running experiments, conducting interviews, administering surveys).
- Store data.
- Document procedures: Keeping detailed records of how the study was conducted, including any deviations from the study protocol and decisions made during execution.

Question: How strongly do you agree that this description appropriately captures the activities involved in the "Data collection" stage of the research cycle?

Consensus: 86%



Summary of comments

Prompt: If you would like to share any changes you would make to the description of the "Data collection" stage of the research cycle, please do so here:

Comments: 24

Summary: While consensus for the description of this stage was high already, eleven panelists commented that data management, including (components of) the FAIR principles could be mentioned more explicitly (see comment 3.1). Five panelists suggested to change the wording,

either so that it is more inclusive of studies that do not collect new data (three panelists; see comment 3.2), or so that it includes other aspects of conducting a study (two panelists). Four panelists noted that "documenting procedures" should be broadened. One panelist specifically noted that documenting procedures not only applies to this stage of the research cycle, but rather to the whole process of doing research (see comment 3.3).

Prompt: If you would like to share any additions you would make to the topics included in the description of the "Data collection" stage of the research cycle, please do so here:

Comments: 24

Summary: Similarly to the previous question, eleven panelists commented that adhering to (components of) the FAIR principles or engaging in data management procedures in general could be added to this stage (see comment 3.4). Five panelists commented that this stage could also include data quality control (see comment 3.5). Four panelists suggested that the documentation of procedures could be broadened. Three of these related to how data collection and data analysis can take place concurrently, meaning that the documentation of procedures applies to both stages.

Selection of comments

Comment 3.1: Storing data also has relevant considerations to it, such as doing it in a safe manner and already thinking about relevant metadata, I miss data management concerns and would expect them here.

Comment 3.2: Again wondering whether retrospective studies are covered (or systematic reviews / meta-analyses by this definition).

Comment 3.3: I don't think "Document procedures: Keeping detailed record of how the study was conducted, including any deviations from the study protocol and decisions made during execution." is actually related to data collection directly. This step actually applies to the whole process, including study planning, data collection, data analysis, and so on?

Comment 3.4: I think that the description needs to further focus on managing the data; making it actually understandable and usable within and outside of the research team. So, managing the data where it is coded properly, has a codebook, etc.

Comment 3.5: Data integrity checks, or safeguards could be considered as part of this step. Many data collection forms contain automated checks, a mention of these might be prudent.

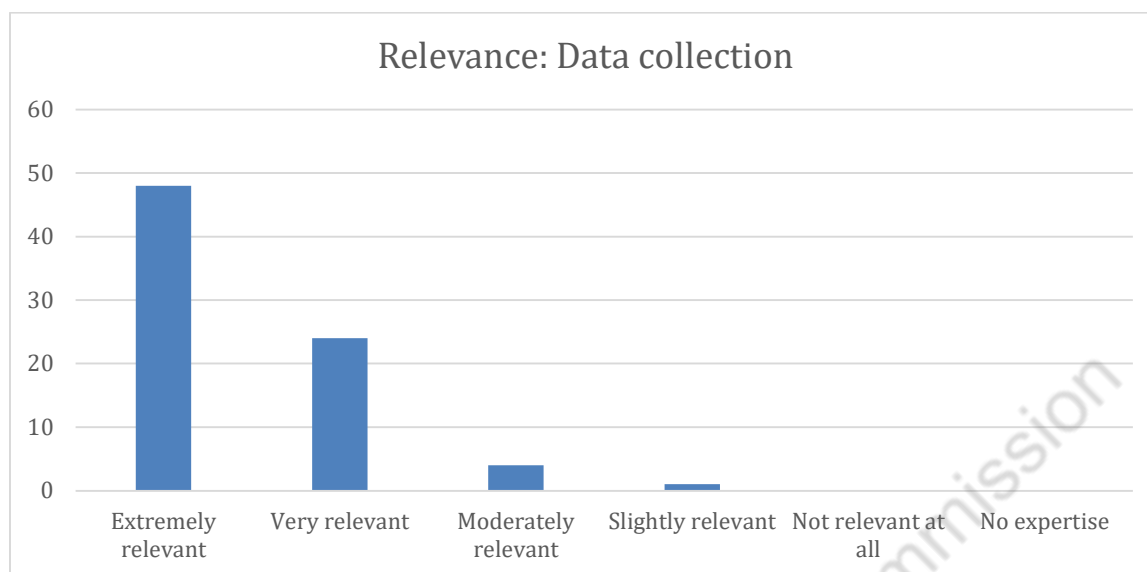
Response of steering committee

While consensus for this description was already substantial (85%), some good suggestions were made by the panelists. Data management and/or FAIR principles could indeed be mentioned more explicitly. Simply "storing data" is not enough. Additionally, we agree that documenting procedures should be applied to the whole process of doing research, and may therefore be more suitable for the category of interventions that spans over the whole research cycle. Lastly, though some panelists rightly noted that the wording of the "collect data" subcategory excludes studies that do not collect new data, we believe that for the purpose of the taxonomy, it is not necessary that every (sub)category applies to every type of research.

Relevance

Question: How relevant is the "Design of the study" stage of the research cycle for facilitating reproducible research practices?

Consensus: 94%



Summary of comments

Prompt: If you would like to share the arguments for your rating of relevance for the "Data collection" stage of the research cycle, please do so here:

Comments: 20

Summary: Ten comments included arguments for why this stage is relevant for facilitating reproducible research practices. Most arguments relate to how documenting (changes in) data collection procedures is important (see comment 3.6). At the same time, one panelist noted that knowing what to document is a barrier here (see comment 3.7). None of the comments argued against the relevance of this stage, though five panelists gave nuanced views, for example stating that it is less relevant than the design (see comment 3.8).

Selection of comments

Comment 3.6: Problems or poor documentation at this phase affect the validity of the data, and therefore the reproducibility of experimental results.

Comment 3.7: This is arguably where many of the problems with reproducibility arise. However, there is a challenge in knowing / noticing and documenting everything that might cause irreproducible research.

Comment 3.8: I rated this as less relevant because if the other researcher does not want to use the original data, only replicate the results, the actual data are less relevant than the design. Also, even though many researchers now routinely share the data, not everyone keeps and shares detailed records.

Response of steering committee

Consensus for the relevance of this stage for facilitating reproducible research practices is high (94%). Based on the comments given by panelists, the most important argument is that poor documentation of data collection procedures can lead to low reproducibility of research results.

5.9.9 Stage 4: Data analysis

Description

Proposed description

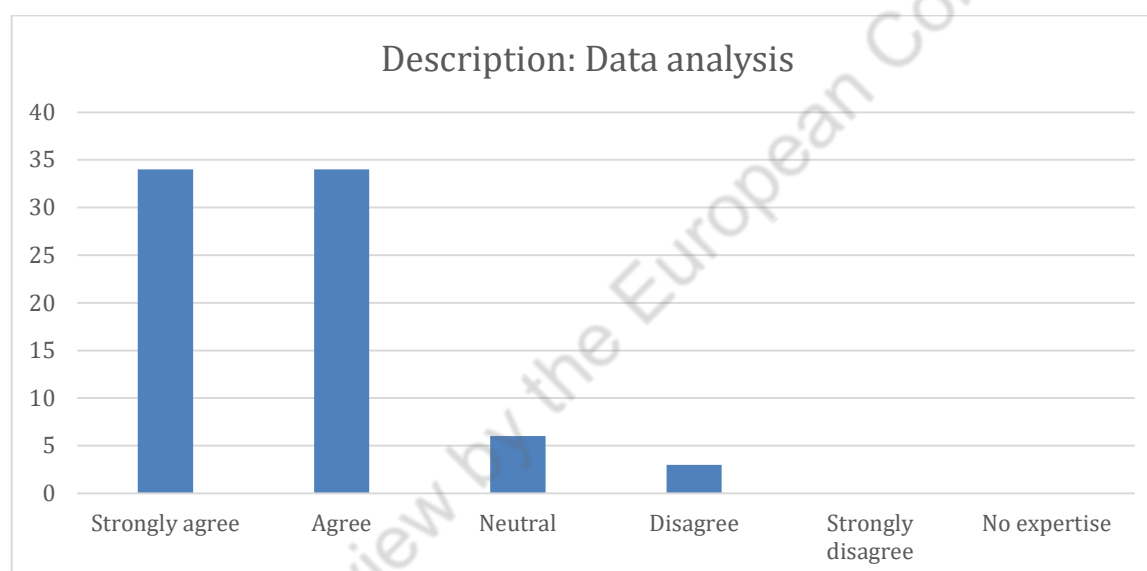
The stage where collected data is prepared, processed, and analyzed to answer the research question.

Includes:

- Prepare and clean data: Organize, transform, and clean raw data to prepare it for analysis.
- Run analyses: Conduct statistical or qualitative analyses.
- Validate results: Assess the robustness and reliability of findings.

Question: How strongly do you agree that this description appropriately captures the activities involved in the "Data analysis" stage of the research cycle?

Consensus: 88%



Summary of comments

Prompt: If you would like to share any changes you would make to the description of the "Data analysis" stage of the research cycle, please do so here:

Comments: 25

Summary: Seven comments proposed a change in wording to improve inclusivity, as the current wording seems to mostly relate to quantitative studies (see comment 4.1). Six panelists commented that assessing the robustness of findings could be named as being optional, as this may not be relevant for any type of research (see comment 4.2). In six comments, panelists commented on the inclusion of data cleaning. Five of these mentioned that data cleaning is a different category than data analysis (see comment 4.3).

Prompt: If you would like to share any additions you would make to the topics included in the description of the "Data analysis" stage of the research cycle, please do so here:

Comments: 21

Summary: Eleven panelists commented that they would add documentation of the analyses, including any deviations from the original plan to this stage of the research cycle (see comment 4.4 and 4.5). Three panelists commented on the importance of data integrity or quality control. Two of these specifically addressed doing assumption checks and analyzing (the impact of) missing data.

Selection of comments

Comment 4.1: “Run analysis” is very quantitative-connotated language. Even just “Analyze the data” is better.

Comment 4.2: Robustness checks are good practice, but they [may] not be warranted/performed in many studies, so once again I think this is not an intrinsic part of the analysis step but an optional one (and perhaps this should be made explicit).

Comment 4.3: It may be better to separate data cleaning & processing from data analysis. These are distinct activities. Data cleaning, preparation & processing can be a lengthy process, and needs to be finished before analysis can start.

Comment 4.4: Due to the importance of the “Data analysis” stage for reproducibility, I believe that a description of data analysis should include a clear description of the preparation of analysis plans and/or code/syntax to enable others to reproduce the same results from the data and through the reported analytical approach.

Comment 4.5: I would suggest adding:

- ‘storing code or log books that describe the previously described three steps in detail’.
- ‘describing any modifications to the original analysis plan’.

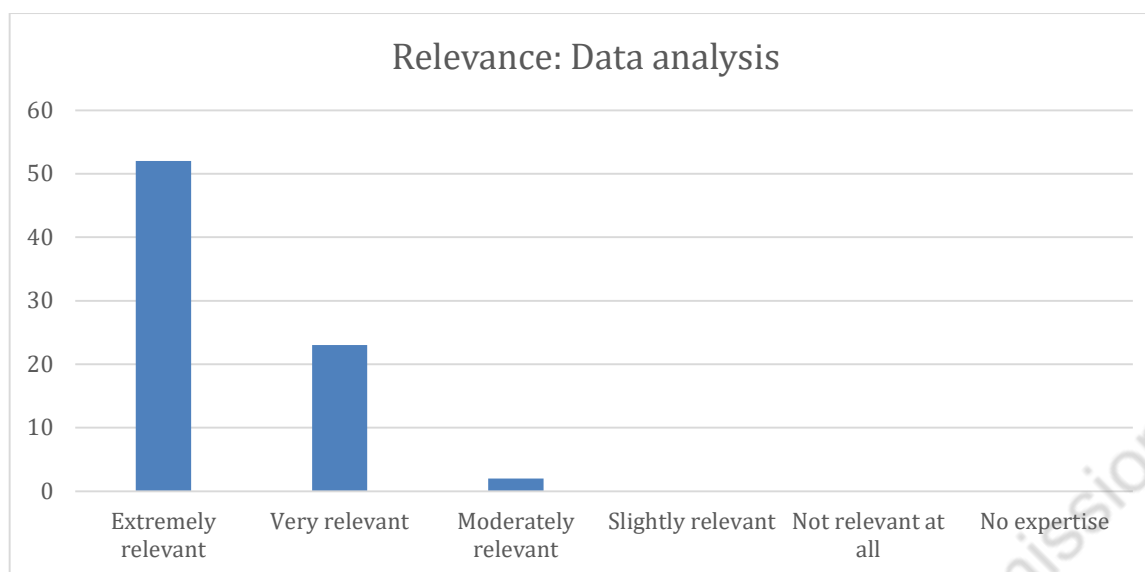
Response of steering committee

Several panelists rightfully commented that “run analyses” seems very quantitative-connotated language, and proposed to change it to improve inclusivity of qualitative data analysis. This point is recognized by the steering committee, and will be implemented in future proposals. Similarly to the previous stage, another rightfully often given comment was that the documentation of procedures is important during this stage of doing research as well.

Relevance

Question: How relevant is the “Data analysis” stage of the research cycle for facilitating reproducible research practices?

Consensus: 97%



Summary of comments

Prompt: If you would like to share the arguments for your rating of relevance for the "Data analysis" stage of the research cycle, please do so here:

Comments: 17

Summary: In ten comments, panelists argue that this stage is relevant for reproducible research practices. Arguments include that data analysis should be done in accordance with the protocol (see comment 4.6), and/or that changes in data analysis should be logged properly. Properly analyzing data is vital for the robustness and trustworthiness of research findings (see comment 4.7). Five comments gave more nuanced views, stating for example that if previous stages are properly done, the data analysis stage may not be as important (see comment 4.8).

Selection of comments

Comment 4.6: Data analysis has to be done carefully and in accordance with the protocols set out in advance. Someone else should be able to reproduce the results if they have the same dataset and apply the same protocol exactly as described.

Comment 4.7: Proper data analysis determines the robustness of results and the reliability of the scientific literature. Since the scientific community builds upon published findings, analyses must ensure accuracy and impartiality to provide trustworthy evidence. Without robust and unbiased results, reproducibility becomes impossible, as one cannot reproduce what never truly occurred.

Comment 4.8: I think in general this stage wouldn't be too impactful for replicability, as (again) most of this should have been established or at least thought of before conducting the study. However, I do see in this stage ample opportunity for data dredging, and other QRPs that may inflate false positives.

Response of steering committee

Consensus on the relevance of this stage for facilitating reproducible research practices was highest of all stages (97%). As mentioned by the panelists, being transparent about (changes

in) data analysis procedures is very important for the reliability of research results. While it is true that a lot of the groundwork for this stage is laid in previous stages such as study design, many QRPs

indeed happen within this stage of the research cycle, making this is a highly relevant stage when it comes to reproducible research practices.

5.9.10 Stage 5: Report the findings

Description

Proposed description

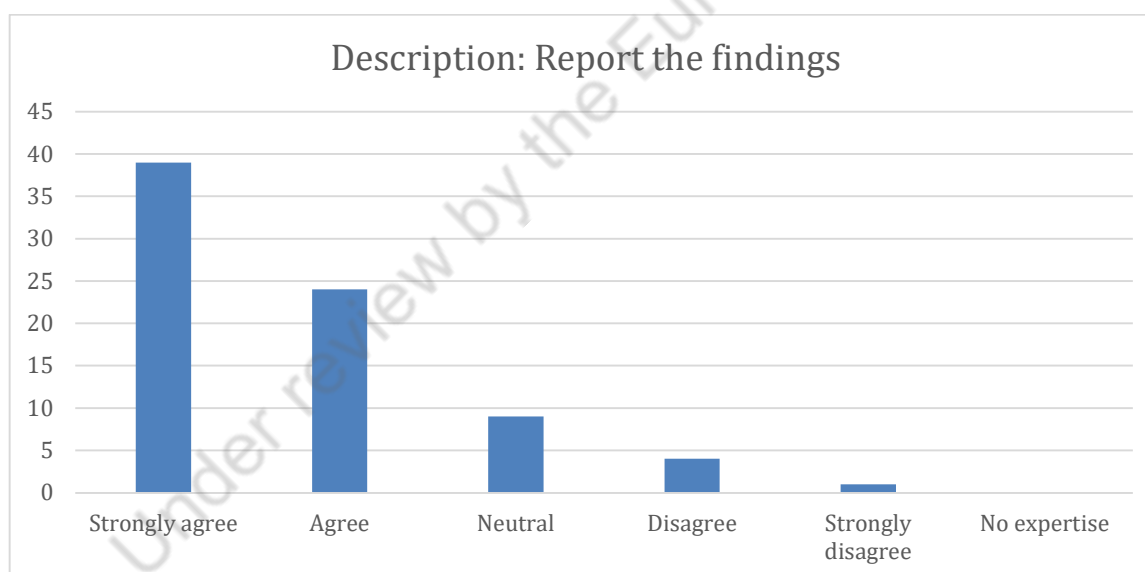
The stage where findings are interpreted and shared with the (scholarly) community.

Includes:

- Interpret findings: Contextualize the results in relation to the research question, hypothesis, theory, and previous literature.
- Write report: Report the findings of the study.
- Publish/disseminate findings: Submit the manuscript to a preprint server, journal or conference.
- Share data/code/materials: Make datasets, analysis scripts, and research materials available for accountability or reuse purposes.

Question: How strongly do you agree that this description appropriately captures the activities involved in the "Report the findings" stage of the research cycle?

Consensus: 82%



Summary of comments

Prompt: If you would like to share any changes you would make to the description of the "Report the findings" stage of the research cycle, please do so here:

Comments: 22

Summary: Six panelists commented on the inclusion of "interpret findings" in this stage. Some note that interpretation of findings better fits with the previous stage (data analysis) or a separate stage (see comment 5.1). Others note that interpreting the findings could be included here, but should then

also be included in the header of this stage (i.e., “Interpret and report the findings”; see comment 5.2). One panelist commented that we interpret data throughout the whole research cycle, not only when we report (see comment 5.3). Seven panelists suggested to include a more broad interpretation of the dissemination of findings, rather than just to preprint servers, journals, or conferences (see comment 5.4). In three comments, panelists

suggested to add a reference to field specific reporting guidelines (see comment 5.5). Five panelists related to components of the FAIR principles in making data available, referring for example how data should be made as open as possible, but as closed as necessary.

Prompt: If you would like to share any additions you would make to the topics included in the description of the "Report the findings" stage of the research cycle, please do so here:

Comments: 20

Summary: Eleven panelists suggested additions that referred to the completeness of reporting, referring for example to the reporting of conflicts of interests, complete and detailed reporting of data collection and analysis method, or complete and unbiased reporting. In ten comments, panelists suggested that data should be made available in a way that is reusable, referring both to accessibility of data (see comment 5.6), as well as completeness of data (see comment 5.7). Three comments included a suggestion for the dissemination of knowledge to a wider public than just the scholarly community.

Selection of comments

Comment 5.1: I would not conflate “interpretation” and “reporting”, I think these are two separate things. Interpretation would be best characterized either as a step in itself or as part of the analysis process. I’d leave a full item just on reporting without the interpretation part.

Comment 5.2: If you include “Interpret findings: Contextualize the results in relation to the research question, hypothesis, theory, and previous literature.” I think the practice should be called “Interpret and report findings”? (or something similar)

Comment 5.3: I think we interpret (and try to make sense of) the data throughout the whole research cycle, not only when we report.

Comment 5.4: I would recommend expanding the description to emphasize strategic dissemination, accessibility, and knowledge translation. Beyond traditional reporting, this stage should highlight the importance of tailoring communication to diverse audiences academic, professional, and community-based ensuring that findings are not only shared but also understood and applied.

Comment 5.5: Add another bullet to recommend using a reporting guideline to report the research.

Comment 5.6: The availability of datasets, analysis scripts, and research materials is a thorny topic – should these also be licensed with pre-specified IP rights and a preservation plan (i.e., stable DOI), or can people hide behind “contact the corresponding author” to which they will never respond when requested. I recognize taxonomically it may be hard to be specific about the method in which these are made available, but I think this is an important thing to consider. What is on paper reported as available by contact may not be truly available in the open science sense of the word.

Comment 5.7: I’m missing a topic about complete and detailed reporting of data collection and analysis method. It is not enough to publish a vague description of the study and provide a plethora of additional pre-processed materials.

Response of steering committee

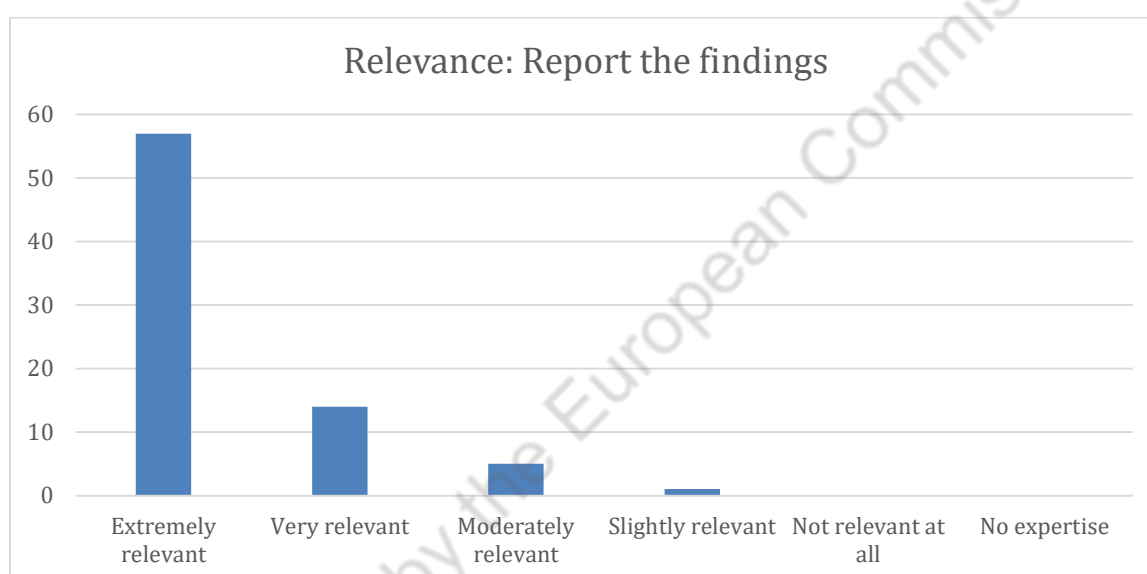
Several panelists commented on the inclusion of “interpret the findings” here. This could reflect the difference between more qualitative approaches and more quantitative ones. For many qualitative researchers the analysis phase is intimately connected to the writing phase - huge parts of the analysis are done while (and through) writing. For most quantitative

researchers, it is easier to separate these two activities. Another commonly given comment related to the completeness of reporting, and the dissemination of findings. All of these are valid suggestions.

Relevance

Question: How relevant is the "Report the findings" stage of the research cycle for facilitating reproducible research practices?

Consensus: 92%



Summary of comments

Prompt: If you would like to share the arguments for your rating of relevance for the "Report the findings" stage of the research cycle, please do so here:

Comments: 24

Summary: Fourteen comments include arguments for why this stage is relevant for facilitating reproducible research practices. Nearly all of these argue along the lines of “without reporting, there is nothing to reproduce” (see comment 5.8). Two panelists raised similar points about the “interpretation of findings”, which was included in the description of this stage (see comment 5.9). Two panelists gave arguments that could be interpreted as arguments against the importance of this stage. One of these panelists mentioned that sharing is important, but not necessarily for reproducibility, while the other commented that you are just executing the plan you set before during this stage. One panelist shared doubts about the importance of this stage (see comment 5.10).

Selection of comments

Comment 5.8: This phase is crucial – one can't distinguish between well designed and poorly designed studies if the research is poorly reported. Reproducibility is impossible if materials need to reproduce or replicate the results are missing.

Comment 5.9: I was in doubt when I first thought about the first bullet point 'interpret the findings'. In principle, this is something that does not have to be reproduced per se; as researchers we can have different interpretations from the same findings. After consideration, I do think it is important though, because a) interpretations are important with respect to how results/evidence are used in future research or practice and b) as a consequence it is important to be transparent in your argumentation. The rest of the bullet points directly relate

to transparency and are thus very relevant for reproducibility.

Comment 5.10: I'm really having doubts about my assessment here. I think the reporting stage may have more relevance for allowing replications to be conducted and assessing reproducibility rather than improving the reproducibility of the results obtained.

Response of steering committee

The vast majority of panelists agreed that this stage is extremely relevant for facilitating reproducible research practices. As mentioned by the panelists, without reporting it is near impossible to distinguish between well and poorly designed studies, and without sharing data (in the broad sense of the word), transparency and reproducibility are at stake.

5.9.11 Stage 6: Post-publication

Description

Proposed description

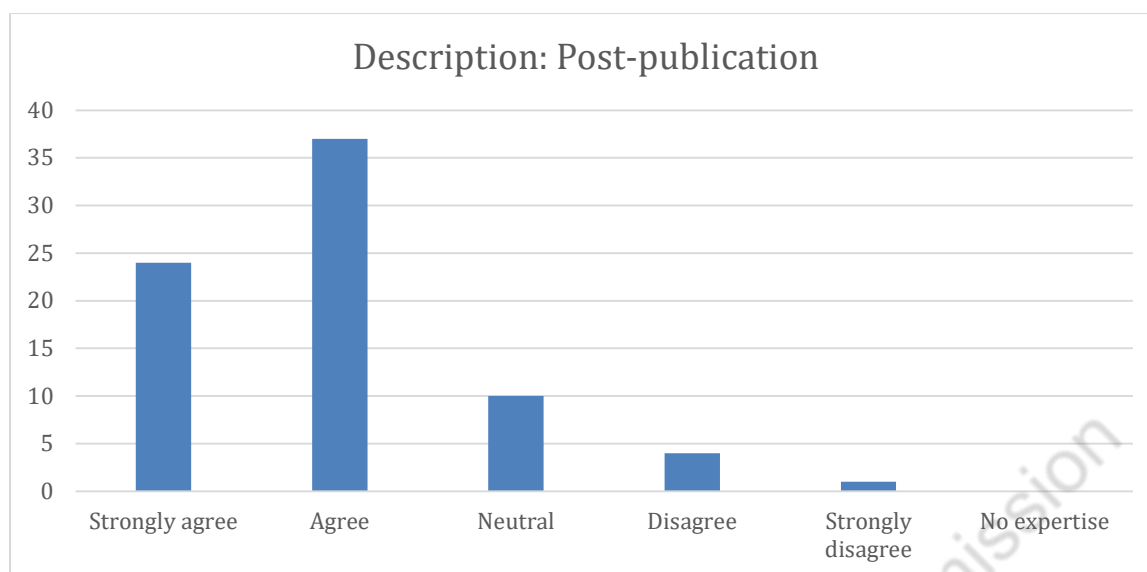
The post-publication stage in which findings are assessed, scrutinized, and extended.

Includes:

- Peer-review and commentary: Participating in or responding to peer review processes that critically assess the study's design, analysis, and interpretation.
- Corrections: Addressing errors, omissions, or limitations discovered after publication by issuing corrections, retractions, or clarifications.
- Replication: Conducting independent attempts to reproduce the findings of the study using the same or closely related methods and materials.

Question: How strongly do you agree that this description appropriately captures the activities involved in the "Post-publication" stage of the research cycle?

Consensus: 79%



Summary of comments

Prompt: If you would like to share any changes you would make to the description of the "Post-publication" stage of the research cycle, please do so here:

Comments: 28

Summary: Ten panelists commented that peer-review mostly happens pre-publication, rather than post-publication (see comment 6.1). Nine panelists noted that replication does not necessarily fit here. They either write that it is not a task of the main authors (see comment 6.2), or – relatedly – that it starts a new research cycle (see comment 6.3).

Prompt: If you would like to share any additions you would make to the topics included in the description of the "Post-publication" stage of the research cycle, please do so here:

Comments: 18

Summary: Four panelists commented that the facilitation of reuse or extension of findings could be added (see comment 6.4). Another four panelists wrote on dissemination in their comment, with three specifically noting that scientific findings should not only be disseminated to the scientific community, but also to society (see comment 6.5). In seven comments, suggestions for alternative categories were given, ranging from building on or extending the research to knowledge mobilization and outreach, computational reproducibility checks, and tracking the reuse of data.

Selection of comments

Comment 6.1: I'd be explicit in describing peer review as "post publication" – most people understand peer review as what happens prior to publication.

Comment 6.2: Conducting independent attempts to reproduce the research is no longer with the main authors of the research (whereas the previous steps were with the main author of the research).

Comment 6.3: I'm not sure whether I'd include replication as a necessary part of the post-publication process – you could also conceptualize it as being a new research cycle.

Comment 6.4: I'd add something like "Facilitate reanalysis/replication/extension of the findings".

Comment 6.5: Scholarly and Public Engagement – Encourage active dialogue with readers, researchers, and community stakeholders through conferences, webinars, blogs, or social media to promote ongoing knowledge exchange.

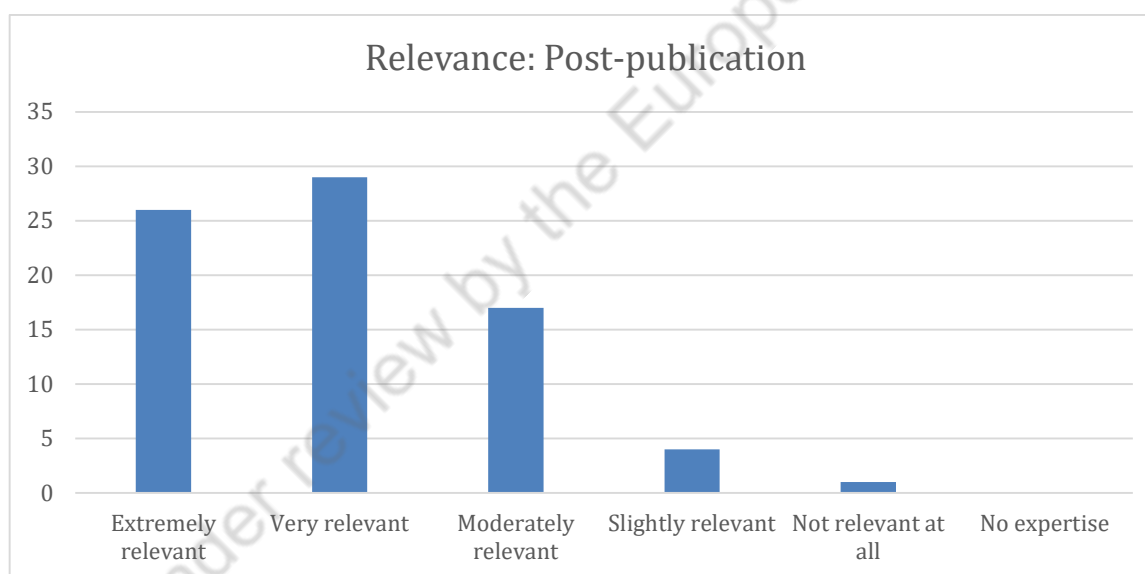
Response of steering committee

Several panelists commented that this stage seems to be "the odd one out", as this stage does not necessarily describe actions that fit within the "regular" research cycle. Rather, these actions start a new research cycle, or are done by different researchers than the authors of the original research. We agree that some of the actions described in this stage are outside the scope of the regular research cycle. At the same time, there are a number of interventions that aim to improve reproducibility that fall within the category of "post-publication", meaning that this category should be included within the taxonomy. As rightly pointed out by multiple panelists, "peer-review" usually refers to a process that happens pre-publication, although post-publication peer-review processes do happen.

Relevance

Question: How relevant is the "Post-publication" stage of the research cycle for facilitating reproducible research practices?

Consensus: 71%



Summary of comments

Prompt: If you would like to share the arguments for your rating of relevance for the "Post-publication" stage of the research cycle, please do so here:

Comments: 19

Summary: Seven comments include arguments in favor of the relevance of this stage for facilitating reproducible research practices. Most of these relate to how this stage is where reproducibility is put into practice (see comment 6.6). Seven panelists gave nuanced answers, mostly by saying that they believe some of the subcategories to be important, but others to be less important (see comment

6.7). Three panelists commented that not every study should be or will be replicated (see comment 6.8).

Selection of comments

Comment 6.6: Very relevant because without replication attempt and without the interactions required at this stage, the term Reproducibility can exist in the void: i.e., one can claim reproducibility but it is at this stage that others are going to attempt replication and further discussion can feed the research that has been conducted.

Comment 6.7: I think replications are extremely important. The other two bullet points should not be important if all things are in order (a detailed pre-registration, adequate study conduct and reporting, sharing of data and code where possible, etc.).

Comment 6.8: It is important to follow the corrections, comments, retractions or whatever it is needed, but having to replicate all your studies is nonsense, time consuming and money constrained. Publishing something that is replicated is still not as widely recognized as regular publication and pressure to publish 'relevant' research is still an ongoing issue, so...

Response of steering committee

Interventions in this stage of the research cycle seem highly relevant for promoting reproducible research practices, as, like some of the panelists mentioned, it is interventions like these that can help establish whether reproducibility is actually promoted through any of the previous actions. At the same time, it also seems true that actions within this phase (such as replication) are not always necessary or preferred.

5.10 Delphi round 2 survey

TRUSTparency Delphi Survey - Round 2 - Reproducibility Experts

Section A: Introduction

Welcome back! This is round 2 of the Delphi survey on reproducibility practices. Process of the Delphi In the next pages, we will ask your opinion on several questions. These questions can be answered on a 5-point scale ranging from "1. Strongly disagree" to "5. Strongly agree". If you feel like you do not have the right expertise to answer a question, you can use the "No expertise" option. After each question we ask you to formulate the rationale behind your answer in an open text field. All open text fields are optional. However, your arguments, ideas, suggestions and feedback are highly valuable to further improve the taxonomy and address our study aims. Please do not include any personally identifying information about people or institutions. All personally identifying information will be deleted from our data. Consensus We will consider to have reached consensus for a question when 67% or more of the panelists have scored it 4 or 5. Based on the argumentation and rate of consensus among the panelists, the steering committee, consisting of the researchers from Amsterdam UMC, will modify, include new, or exclude elements and/or their descriptions. This will be presented in the next round and the steering committee will communicate all decisions made. Notes on terminology Before we start, we first provide you with some notes on terminology, as the concept of reproducibility remains ambiguously and inconsistently defined across research fields. When we talk about reproducible research practices, we refer to practices that enhance the reproducibility of research. When we talk about reproducibility, we refer to the extent to which the results of a replication agree with those of the original study. Lastly, when we talk about replication, we refer to an actual attempt to re-produce an earlier finding or experiment. Let's get started! On the following pages, we will ask you questions regarding the taxonomy, relating back to the first round. Compared to the first round, we made three large changes. We will start with these changes. For each proposed change, we ask you to tell us the extent to which you agree with this proposal. We also ask you to formulate the rationale for your answers. After these three main changes, we will give you the revised descriptions for the subcategories, based on the input that we received in the first round. Again, we will ask you to tell us the extent to which you agree with each description. At the end of the survey, we will ask you if you want to share any final reflections.

Page 1

Section B: The main changes

We will start with the main changes that were made based on the comments we received in the first round of this Delphi survey.

B1. The first main change In the first round, we presented the following six stages: Formulate a research question/hypothesis Design of the study Data collection Data analysis Report the findings Post-publication Consensus rates for the first and last stage that we presented were relatively low compared to the four middle stages (71% and 79%, respectively, compared to 82-88% for the description, and 71% and 71% compared to 92-97% for relevance). Many panelists commented that they were unsure about the inclusion of "obtain resources" in the first stage of the research cycle. Some panelists noted that the first and the second stage seemed to overlap. As both the first (Formulate a research question/hypothesis) and the second stage (Design of the study) that we presented relate to the preparation of a study, we propose to merge these into one category: preparation. To what extent do you agree with the proposal to merge "Formulate a research question/hypothesis" and "Design of the study" into one category: "Preparation"?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

B2. If you would like to share the arguments for your rating of agreement, please do so here:

Page 2

LimeSurvey

B3. The second main change Based on the comments given in the first round of the Delphi, we realized that there is a category of interventions that have a different status than interventions that are implemented throughout the research cycle by researchers themselves. These refer to interventions that can be implemented in a foundational, pre-research cycle stage. Examples of these would be interventions implemented by funders, such as mandates for writing a data management plan or for making data openly available. Interventions as such form the infrastructure that facilitate the engagement of reproducible research practices by researchers. We therefore propose to add an "Infrastructure" category, intended for those interventions that form the infrastructure for reproducibility. Subcategories of this category will be "Research performing organizations (RPOs)", "Research Funding Organizations (RFOs)", and "Publishers". To what extent do you agree with the proposal to add an "Infrastructure" category?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

B4. If you would like to share the arguments for your rating of agreement, please do so here:

Page 3

LimeSurvey

B5. The third main change Third, consensus for the post-publication stage was slightly lower than for other stages (i.e., 79%). While this is still high, we received valid comments from panelists who said that peer-review happens mostly pre-publication rather than post-publication, and that these two should be distinguished from each other. Based on the comments that were given, we propose to re-name the fifth stage from "Write the report" to "Disseminate", and to make "Post-publication" (originally the 6th stage) and "Peer-review" into subcategories of this stage. All categories within this stage will include interventions that are implemented after data has been collected and analyzed, and is ready to be disseminated. To what extent do you agree with the proposal to re-name the "Write the report" stage to "Disseminate", and to add "Post-publication" and "Peer-review" as subcategories to this stage?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

B6. If you would like to share the arguments for your rating of agreement, please do so here:

Section C: That concludes the main changes.
You have now answered all questions regarding the main changes that we propose based on the comments that we received in round 1 of the Delphi survey. We will now present the descriptions of the subcategories for each stage and ask for your consensus on the next pages. Note that the examples given with the descriptions are not exhaustive. The goal of the taxonomy is to be able to categorize all interventions that aim to improve reproducibility into the categories that we present. That means that the descriptions of the categories are relatively broad, so that different interventions can fit into the category. This is an overview of the full proposed taxonomy:

Page 4

LimeSurvey

Section D: Stage: Preparation
 Description: The stage in which the study is conceptually and practically planned and prepared. Includes: Formulate the research question/hypothesis; the process of defining what the study aims to investigate; Plan the study: creating a plan for how the study will be conducted; Obtain resources: securing the resources necessary to execute the study.

D1. The first subcategory for this stage is: Formulate the research question/hypothesis. Description: the process of defining what the study aims to investigate.
 Examples of interventions that could fall into this category are: conducting literature reviews; using reporting guidelines or checklists for formulating research questions/hypotheses; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

D2. If you would like to share the arguments for your rating of agreement, please do so here:

D3. The second subcategory for this stage is: Plan the study. Description: creating a plan for how the study will be conducted.
 Examples of interventions that could fall into this category are: developing a study protocol; writing a data management plan; writing a statistical analysis plan; preregistration; pilot testing; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

LimeSurvey

D4. If you would like to share the arguments for your rating of agreement, please do so here:

D5. The third subcategory for this stage is: Obtaining resources. Description: securing the resources necessary to execute the study.
 Examples of interventions that could fall into this category are: assembling or training the research team; obtaining or creating materials; requesting funding; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

D6. If you would like to share the arguments for your rating of agreement, please do so here:

D7. Are there any subcategories that you are currently missing within the "Preparation" stage? In other words: are there any interventions aimed at enhancing reproducible research practices within this stage of the research cycle that cannot be put into any of the presented subcategories?

LimeSurvey

Section E: Stage: Data collection
 Description: The stage in which data are collected and stored Includes: Data collection: executing the data-gathering process
 Data storage: saving and organizing the collected data Document procedures: recording details of how the data are collected

E1. The first subcategory of this stage is: Data collection
 Description: executing the data-gathering process
 Examples of interventions that could fall into this category are: standardization of data collection procedures; training in data collection methods; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No experience ☐

E2. If you would like to share the arguments for your rating of agreement, please do so here:

E3. The second subcategory of this stage is: Data storage
 Description: saving and organizing the collected data
 Examples of interventions that could fall into this category are: using secure data repositories; use of data management procedures; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No experience ☐

LimeSurvey

E4. If you would like to share the arguments for your rating of agreement, please do so here:

E5. The third subcategory of this stage is: Document procedures
 Description: recording details of how the data are collected
 Examples of interventions that could fall into this category are: maintaining lab notebooks; documenting changes to the original protocol; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No experience ☐

E6. If you would like to share the arguments for your rating of agreement, please do so here:

E7. Are there any subcategories that you are currently missing within the "Data collection" stage? In other words: are there any interventions aimed at enhancing reproducible research practices within this stage of the research cycle that cannot be put into any of the presented subcategories?

LimeSurvey

Section F: Stage: Data analysis

Description: The stage where data are prepared and analyzed to answer the research question. Includes: Data preparation: organizing and cleaning the data to prepare for analysis; Conduct analyses: conducting quantitative and/or qualitative analyses; Data validation: assessing the accuracy and reliability of findings; Document procedures: recording details of how the data are analyzed.

F1. The first subcategory of this stage is: Data preparation. Description: organizing and cleaning the data to prepare for analysis. Examples of interventions that could fall into this category are: converting data into interoperable formats; using controlled vocabularies or coding schemes to standardize variables; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

F2. If you would like to share the arguments for your rating of agreement, please do so here:

F3. The second subcategory of this stage is: Conduct analyses. Description: conducting quantitative and/or qualitative analysis. Examples of interventions that could fall into this category are: use of open source analysis software; applying standardized analysis methods; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

LimeSurvey

F4. If you would like to share the arguments for your rating of agreement, please do so here:

F5. The third subcategory of this stage is: Data validation. Description: assessing the accuracy and reliability of findings. Examples of interventions that could fall into this category are: conducting robustness checks; independent re-analysis of results; data integrity checks; using an effect-size filter; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

F6. If you would like to share the arguments for your rating of agreement, please do so here:

LimeSurvey

F7. The fourth subcategory of this stage is: Document procedures
Description: recording details of how the data are analyzed
Examples of interventions that could fall into this category are: maintaining an analysis log or notebook; using version controlling for data analysis; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

F8. If you would like to share the arguments for your rating of agreement, please do so here:

F9. Are there any subcategories that you are currently missing within the "Data analysis" stage? In other words: are there any interventions aimed at enhancing reproducible research practices within this stage of the research cycle that cannot be put into any of the presented subcategories?

LimeSurvey

Section G: Stage: Dissemination
Description: The stage where findings and data are communicated and shared with the scientific community and the public. Includes: Write the report: reporting the findings of the study Share data: making research data accessible to others Peer-review: the evaluation of research by others Post-publication: contributing to the ongoing use, evaluation, and discussion of research after publication

G1. The first subcategory of this stage is: Write the report
Description: report the findings of the study
Examples of interventions that could fall into this category are: complete reporting; using a reporting guideline; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

G2. If you would like to share the arguments for your rating of agreement, please do so here:

G3. The second subcategory of this stage is: Share data
Description: making research data accessible to others
Examples of interventions that could fall into this category are: sharing data in repositories; applying the FAIR-principles; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

LimeSurvey

G4. If you would like to share the arguments for your rating of agreement, please do so here:

G5. The third subcategory of this stage is: Peer-reviewDescription: the evaluation of research by others
Examples of interventions that could fall into this category are: use of reproducibility checklist during peer-review; guideline reminders for peer-reviewers; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

G6. If you would like to share the arguments for your rating of agreement, please do so here:

Page 13

LimeSurvey

G7. The fourth subcategory of this stage is: Post-publicationDescription: contributing to the ongoing use, evaluation, and discussion of research after publication
Examples of interventions that could fall into this category are: writing letters to editors or authors to highlight issues; engaging in replication studies; responding to requests to re-use data; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐

2. Disagree ☐

3. Neutral ☐

4. Agree ☐

5. Strongly agree ☐

No expertise ☐

G8. If you would like to share the arguments for your rating of agreement, please do so here:

G9. Are there any subcategories that you are currently missing within the "Dissemination" stage? In other words: are there any interventions aimed at enhancing reproducible research practices within this stage of the research cycle that cannot be put into any of the proposed subcategories?

Page 14

LimeSurvey

Section H: Category: Infrastructure
 Description: The systems, policies, and support provided by organizations to enable and promote reproducible research practices throughout the research cycle. Includes: Research performing organizations (RPOs): Organizations that perform research and provide resources, policies, and support for their researchers. Research funding organizations (RFOs): Organizations that fund research and set requirements or incentives for reproducibility. Publishers: Organizations that disseminate research outputs and set standards for transparency and reproducibility in publication.

H1. The first subcategory of this category is: **Research Performing Organizations (RPOs)** Description: organizations that perform research and provide resources, policies, and support for their researchers.
 Examples of interventions that could fall into this category are: offering training and support in open science and/or reproducible research; implementing policies on data management; recognizing and rewarding reproducible research practices in hiring and promotion; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

H2. If you would like to share the arguments for your rating of agreement, please do so here:

Page 15

LimeSurvey

H3. The second subcategory of this category is: **Research Funding Organizations (RFOs)** Description: organizations that fund research and set requirements or incentives for reproducibility.
 Examples of interventions that could fall into this category are: requiring data management plans and/or open data sharing as part of grant applications; funding infrastructure for open science/reproducibility; mandating preregistration for funded projects; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

H4. If you would like to share the arguments for your rating of agreement, please do so here:

H5. The third subcategory of this category is: **Publishers** Description: organizations that disseminate research and set standards for transparency and reproducibility in publication.
 Examples of interventions that could fall into this category are: requiring data and/or code availability statements in publications; offering registered reports options; implementing reproducibility checklists; recognizing and rewarding reproducible research practices through badges; etc. To what extent do you agree with the description of this subcategory?

1. Strongly disagree ☐
 2. Disagree ☐
 3. Neutral ☐
 4. Agree ☐
 5. Strongly agree ☐
 No expertise ☐

Page 16

LimeSurvey

H6. If you would like to share the arguments for your rating of agreement, please do so here:

H7. Are there any subcategories that you are currently missing within the "Infrastructure" category? In other words: are there any interventions aimed at enhancing reproducible research practices within this category that cannot be put into any of the proposed subcategories?

Section I: That concludes the descriptions of the subcategories of the taxonomy. If there is anything else you would like to share, please do so here:
This is an overview of the full proposed taxonomy:

I1.

LimeSurvey

Section J: This is the end of round 2!

J1.

In a few weeks, we will contact you again to distribute the feedback report for this round of the Delphi study. If necessary, we may also set up a third round. In case you do NOT want to be contacted again, please tick the option below. If you are happy to be contacted again, you may skip this question.

I do NOT want to be contacted about this study again ☐

Thank you for your participation! This is the end of the survey. Unless you indicated that you no longer want to be contacted regarding this Delphi study, we may contact you within the next few weeks in case there will be a third round of the survey.

5.11 Reference numbers to the combined results of the TIER2/OSIRIS and TRUSTparency scoping reviews

This is a combined list of the publications that were the results of the TIER2/OSIRIS and TRUSTparency projects. The numbering of these publications are used as a reference in Table 16 of section 3.3.3.

Reference No.	Source	First author surname	Publication title
1	TRUSTparency	Abdelaziz (2023)	Optimizing large real-world data analysis with parquet files in R: A step-by-step tutorial
2	TRUSTparency	Aguilar Gomez (2023)	FAIR EVA: Bringing institutional multidisciplinary repositories into the FAIR picture
3	TRUSTparency	Ahern (2024)	Adjusting Adjustments Using External Data to Estimate the Impact of Different Confounder Sets on Published Associations
4	TRUSTparency	Anido-Alonso (2023)	Decentralized data-privacy preserving deep-learning approaches for enhancing inter-database generalization in automatic sleep staging
5	TRUSTparency	Ashkarran (2024)	Standardizing Protein Corona Characterization in Nanomedicine: A Multicenter Study to Enhance Reproducibility and Data Homogeneity
6	TRUSTparency	Bahl (2023)	PROTEOMAS: a workflow enabling harmonized proteomic meta-analysis and proteomic signature mapping
7	TRUSTparency	Bontempi (2024)	End-to-end reproducible AI pipelines in radiology using the cloud
8	TRUSTparency	Bramley (2024)	Ask, and it shall be given you – individual patient data and code availability for randomised controlled trials submitted for publication.
9	TRUSTparency	Chen (2025)	Addressing Clinical Challenges in Aberrant Pharmacokinetics of Biologic Therapeutic Drugs: Investigating Sample Processing Procedure in the Immunoassays
10	TRUSTparency	Denisova (2024)	Enabling reliable usability assessment and comparative analysis of medical software: a comprehensive framework for multimodal biomedical imaging platforms
11	TRUSTparency	De Vries (2024)	ECOTOXr: An R package for reproducible and transparent retrieval of data from EPA's ECOTOX database
12	TRUSTparency	Ellis (2024)	Assessing computational reproducibility in Behavior Research Methods
13	TRUSTparency	Elmanzalawi (2025)	DFAST_QC: quality assessment and taxonomic identification tool for prokaryotic Genomes
14	TRUSTparency	Fernholz (2024)	DeepD3, an open framework for automated quantification of dendritic spines
15	TRUSTparency	Ferreira (2023)	I Choose You: Selecting Accurate Reference Genes for qPCR Expression Analysis in Reproductive Tissues in Arabidopsis thaliana
16	TRUSTparency	Fiedler (2023)	Standardized tissue sampling guidelines for histopathological and molecular analyses of rainbow trout (<i>Oncorhynchus mykiss</i>) in ecotoxicological studies
17	TRUSTparency	Filatovas (2024)	Advancing Research Reproducibility in Machine Learning through Blockchain Technology
18	TRUSTparency	Fraser (2023)	Predicting reliability through structured expert elicitation with the replicATS (Collaborative Assessments for Trustworthy Science) process
19	TRUSTparency	Gaignard (2023)	FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards
20	TRUSTparency	Gierend (2023)	Traceable Research Data Sharing in a German Medical Data Integration Center With FAIR (Findability, Accessibility, Interoperability, and Reusability)-Geared Provenance Implementation: Proof-of-Concept Study
21	TRUSTparency	Gillies (2024)	Medicine utilization studies in Australian individual-level dispensing data: A blinded, multi-center replicated analysis
22	TRUSTparency	Goetze (2024)	Simultaneous targeted and discovery-driven clinical proteotyping using hybrid-PRM/DIA
23	TRUSTparency	Hall (2023)	Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components
24	TRUSTparency	Haven (2023)	Biomedical supervisors' role modeling of open science practices

25	TRUSTparenity	Hayashi (2023)	brainlife.io: a decentralized and open-source cloud platform to support neuroscience research
26	TRUSTparenity	Hellen (2023)	LiverQuant: An Improved Method for Quantitative Analysis of Liver Pathology
27	TRUSTparenity	Huckvale (2023)	Benchmark Dataset for Training Machine Learning Models to Predict the Pathway Involvement of Metabolites
28	TRUSTparenity	Jaric (2024)	Using mice from different breeding sites fails to improve replicability of results from single-laboratory studies
29	TRUSTparenity	Kalvaityte (2024)	Rapid generation of single-insertion transgenics by Tol2 transposition in zebrafish
30	TRUSTparenity	Kang (2023)	Study design features that improve effect sizes in cross-sectional and longitudinal brain-wide association studies
31	TRUSTparenity	Kekecs (2023)	Raising the value of research studies in psychological science by increasing the credibility of research reports: the transparent Psi project
32	TRUSTparenity	Kessel (2024)	Promoting open science in test-driven software experiments
33	TRUSTparenity	Kolokathis (2024)	Easy-MODA: Simplifying standardised registration of scientific simulation workflows through MODA template guidelines powered by the Enalos Cloud Platform
34	TRUSTparenity	Krahmer (2023)	Care to share? Experimental evidence on code sharing behavior in the social sciences
35	TRUSTparenity	Landau (2024)	Positron emission tomography harmonization in the Alzheimer's Disease Neuroimaging Initiative: A scalable and rigorous approach to multisite amyloid and tau quantification
36	TRUSTparenity	Liu (2024)	Challenging the N-Heuristic: Effect size, not sample size, predicts the replicability of psychological science
37	TRUSTparenity	Li (2024)	ReUseData: an R/Bioconductor tool for reusable and reproducible genomic data management
38	TRUSTparenity	Llovera (2024)	Stroke of Consistency: Streamlining Multicenter Protocols for Enhanced Reproducibility of Infarct Volumes in Preclinical Stroke Research
39	TRUSTparenity	Luna (2024)	Digital RNA sequencing using unique molecular identifiers enables ultrasensitive RNA mutation analysis
40	TRUSTparenity	Lusa (2024)	Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis
41	TRUSTparenity	Mager (2023)	Interlaboratory Reproducibility in Growth and Reporter Expression in the Cyanobacterium Synechocystis sp. PCC 6803
42	TRUSTparenity	Martinc (2024)	Textflows: an open science NLP evaluation approach
43	TRUSTparenity	Martin (2023)	Enhancing collaborative neuroimaging research: introducing COINSTAC Vaults for federated analysis and reproducibility
44	TRUSTparenity	Miyazaki (2023)	Crystallization of amorphous nifedipine under isothermal conditions: Inter-laboratory reproducibility and investigation of the factors affecting reproducibility
45	TRUSTparenity	Nance (2024)	The Causal Roadmap and Simulations to Improve the Rigor and Reproducibility of Real-data Applications
46	TRUSTparenity	Omondi (2024)	Improving rigor and reproducibility in western blot experiments with the blotRig analysis
47	TRUSTparenity	Payares-Garcia (2023)	NeuroNorm: An R package to standardize multiple structural MRI
48	TRUSTparenity	Pelletier (2024)	BERNN: Enhancing classification of Liquid Chromatography Mass Spectrometry data with batch effect removal neural networks
49	TRUSTparenity	Rehal (2023)	Handling intercurrent events and missing data in non-inferiority trials using the estimand framework: A tuberculosis case study
50	TRUSTparenity	Rengasamy (2023)	Exploring the multiverse: the impact of researchers' analytic decisions on relationships between depression and inflammatory markers
51	TRUSTparenity	Rieger (2024)	LDAPprototype: a model selection algorithm to improve reliability of latent Dirichlet allocation
52	TRUSTparenity	Rojas-Velazquez (2024)	Methodology for biomarker discovery with reproducibility in microbiome data using machine learning
53	TRUSTparenity	Schacherer (2023)	The NCI Imaging Data Commons as a platform for reproducible research in computational pathology
54	TRUSTparenity	Schuetz (2023)	vexptoolbox: A software toolbox for human behavior studies using the Vizard virtual reality platform

55	TRUSTparenecy	Senior (2023)	Estimating Genetic Variance in Life-Span Response to Diet: Insights From Statistical Simulation
56	TRUSTparenecy	Souza (2023)	Strategies for consistent and automated quantification of HDL proteome using data-independent acquisition
57	TRUSTparenecy	Sprang (2024)	HDL proteome using data-independent acquisition
58	TRUSTparenecy	Stahl (2023)	Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in European Radiology?
59	TRUSTparenecy	Tang (2024)	EigenRF: an improved metabolomics normalization method with scores for reproducibility evaluation on importance rankings of differential metabolites
60	TRUSTparenecy	Tång-Hallbäck (2025)	Evaluation of nationwide analysis surveillance for methicillin-resistant Staphylococcus aureus within Genomic Medicine Sweden
61	TRUSTparenecy	van Houdt (2023)	Contrast-agent-based perfusion MRI code repository and testing framework: ISMRM Open Science Initiative for Perfusion Imaging (OSIPI)
62	TRUSTparenecy	Van Kampen (2024)	ENCORE: a practical implementation to improve reproducibility and transparency of computational research
63	TRUSTparenecy	Viant (2024)	Demonstrating the reliability of in vivo metabolomics based chemical grouping: towards best practice
64	TRUSTparenecy	Wallace (2024)	Pioneering a multi-phase framework to harmonize self-reported sleep data across cohorts
65	TRUSTparenecy	Wirz (2024)	Increasing the Reproducibility and Replicability of Supervised AI/ML in the Earth Systems Science by Leveraging Social Science Methods
66	TRUSTparenecy	Yousefirizi (2024)	From code sharing to sharing of implementations: Advancing reproducible AI development for medical imaging through federated testing
67	TRUSTparenecy	Zhang (2023)	Highly reproducible and cost-effective one-pot organoid differentiation using a novel platform based on PF-127 triggered spheroid assembly
68	TRUSTparenecy	Zhu (2023)	Documentation strategy for facilitating the reproducibility of geo-simulation experiments
69	TIER2/OSIRIS	Agha (2016)	Impact of the mandatory implementation of reporting guidelines on reporting quality in a surgical journal: A before and after study
70	TIER2/OSIRIS	Agha (2017)	Impact of the PROCESS guideline on the reporting of surgical case series: A before and after study
71	TIER2/OSIRIS	Alsheikh-Ali (2011)	Public availability of published research data in high-impact journals
72	TIER2/OSIRIS	Arroyo-Araujo (2019)	Reproducibility via coordinated standardization: a multi-center study in a Shank2 genetic rat model for Autism Spectrum Disorders
73	TIER2/OSIRIS	Askarov (2022)	The Significance of Data-Sharing Policy
74	TIER2/OSIRIS	Audet (2017)	Examining the adherence to CONSORT guidelines and the reporting of the enrollment process in clinical oncology randomized controlled trials: A review of trials published between 2013 and 2015
75	TIER2/OSIRIS	Auspurg (2020)	More openness in research? An evaluation of open science measures at the Journal of Sociology
76	TIER2/OSIRIS	Barnes (2015)	Impact of an online writing aid tool for writing a randomized trial report: the COBWEB (Consort-based WEB tool) randomized controlled trial
77	TIER2/OSIRIS	Bergeat (2022)	Data Sharing and Reanalyses Among Randomized Clinical Trials Published in Surgical Journals Before and After Adoption of a Data Availability and Reproducibility Policy
78	TIER2/OSIRIS	Bergeat (2022)	Data Sharing and Reanalyses Among Randomized Clinical Trials Published in Surgical Journals Before and After Adoption of a Data Availability and Reproducibility Policy
79	TIER2/OSIRIS	Blanco (2020)	Effect of an editorial intervention to improve the completeness of reporting of randomised trials: a randomised controlled trial
80	TIER2/OSIRIS	Blanco-Perez (2020)	Publication Bias and Editorial Statement on Negative Findings
81	TIER2/OSIRIS	Can (2011)	Has the quality of abstracts for randomised controlled trials improved since the release of Consolidated Standards of Reporting Trial guideline for abstract reporting? A survey of four high-profile anaesthesia journals
82	TIER2/OSIRIS	Caron (2017)	A Survey of the Prevalence and Impact of Reporting Guideline Endorsement in Pathology Journals

83	TIER2/OSIRIS	Caron (2017)	A Survey of the Prevalence and Impact of Reporting Guideline Endorsement in Pathology Journals
84	TIER2/OSIRIS	Chan (2023)	rang: Reconstructing reproducible R computational environments
85	TIER2/OSIRIS	Chow (2018)	The degree of adherence to CONSORT reporting guidelines for the abstracts of randomised clinical trials published in anaesthesia journals: A cross-sectional study of reporting adherence in 2010 and 2016
86	TIER2/OSIRIS	Clayson (2019)	Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability
87	TIER2/OSIRIS	Cobo (2011)	Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial
88	TIER2/OSIRIS	Crüwell (2023)	What's in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in One Issue of Psychological Science
89	TIER2/OSIRIS	Culina (2020)	Low availability of code in ecology: A call for urgent action
90	TIER2/OSIRIS	Danchev (2021)	Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement
91	TIER2/OSIRIS	Deardorff (2020)	Assessing the impact of introductory programming workshops on the computational reproducibility of biomedical workflows
92	TIER2/OSIRIS	Federer (2018)	Data sharing in PLOS ONE: An analysis of Data Availability Statements
93	TIER2/OSIRIS	Fernandes (2019)	Methodological standards, quality of reporting and regulatory compliance in animal research on amyotrophic lateral sclerosis: a systematic review
94	TIER2/OSIRIS	Findley (2016)	Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study
95	TIER2/OSIRIS	Fuller (2014)	Impact of the transparent reporting of evaluations with nonrandomized designs reporting guideline: Ten years on
96	TIER2/OSIRIS	Giofre (2017)	The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices
97	TIER2/OSIRIS	Giofre (2017)	The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices
98	TIER2/OSIRIS	Giofre (2022)	The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices: Five years later
99	TIER2/OSIRIS	Giofre (2022)	The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices: Five years later
100	TIER2/OSIRIS	Gkiouras (2022)	A Meta-Epidemiological Study of Positive Results in Clinical Nutrition Research: The Good, the Bad and the Ugly of Statistically Significant Findings
101	TIER2/OSIRIS	Goldacre (2019)	COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time
102	TIER2/OSIRIS	Goldacre (2019)	COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time
103	TIER2/OSIRIS	Gorman (2020)	Availability of Research Data in High-Impact Addiction Journals with Data Sharing Policies
104	TIER2/OSIRIS	Gregori (2013)	An effect size filter improves the reproducibility in spectral counting-based comparative proteomics
105	TIER2/OSIRIS	Gruber (2016)	Limitations of empirical calibration of p-values using observational data
106	TIER2/OSIRIS	Gruber (2016)	Limitations of empirical calibration of p-values using observational data
107	TIER2/OSIRIS	Hamilton (2022)	How often do cancer researchers make their data and code available and what factors are associated with sharing?
108	TIER2/OSIRIS	Hamilton (2022)	How often do cancer researchers make their data and code available and what factors are associated with sharing?
109	TIER2/OSIRIS	Hamilton (2023)	Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data

110	TIER2/OSIRIS	Hardt (2013)	Surgical trials and trial registers: A cross-sectional study of randomized controlled trials published in journals requiring trial registration in the author instructions
111	TIER2/OSIRIS	Hardt (2013)	Surgical trials and trial registers: A cross-sectional study of randomized controlled trials published in journals requiring trial registration in the author instructions
112	TIER2/OSIRIS	Hardwicke (2018)	Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition
113	TIER2/OSIRIS	Hardwicke (2021)	Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: an observational study
114	TIER2/OSIRIS	Haven (2023)	Biomedical supervisors' role modeling of open science practices
115	TIER2/OSIRIS	Hopewell (2012)	Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: Interrupted time series analysis
116	TIER2/OSIRIS	Hopewell (2016)	Impact of a web-based tool (WebCONSORT) to improve the reporting of randomised trials: results of a randomised controlled trial
117	TIER2/OSIRIS	Howell (2015)	The effect of the SQUIRE (Standards of Quality Improvement Reporting Excellence) guidelines on reporting standards in the quality Improvement literature: A before-and-after study
118	TIER2/OSIRIS	Jaljuli (2023)	A multi-lab experimental assessment reveals that replicability can be improved by using empirical estimates of genotype-by-lab interaction
119	TIER2/OSIRIS	Kidwell (2016)	Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency
120	TIER2/OSIRIS	Krawczyk (2012)	(Un)available upon request: field experiment on researchers' willingness to share supplementary materials
121	TIER2/OSIRIS	Ladd (2023)	An evaluation of reporting guidelines and Clinical Trial Registry requirements among anesthesiology journals
122	TIER2/OSIRIS	Ladd (2023)	An evaluation of reporting guidelines and Clinical Trial Registry requirements among anesthesiology journals
123	TIER2/OSIRIS	Laurinavichyute (2022)	Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy
124	TIER2/OSIRIS	Laurinavichyute (2022)	Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy
125	TIER2/OSIRIS	Liu (2022)	Inadequate awareness of adherence to ARRIVE guidelines, regarding reporting quality of hernia models repaired with meshes: a systematic review
126	TIER2/OSIRIS	Madduri (2019)	Reproducible big data science: A case study in continuous FAIRness
127	TIER2/OSIRIS	Magee (2014)	The dawn of open access to phylogenetic data
128	TIER2/OSIRIS	Marin (2015)	FDAAA legislation is working, but methodological flaws undermine the reliability of clinical trials: a cross-sectional study
129	TIER2/OSIRIS	McGee (2011)	Trial registration and declaration of registration by authors of randomized controlled trials
130	TIER2/OSIRIS	Menon (2021)	The impact of conducting preclinical systematic reviews on researchers and their research: A mixed method case study
131	TIER2/OSIRIS	Meredith (2019)	Biospecimen Data Reporting in the Research Literature
132	TIER2/OSIRIS	Milia (2012)	Mine, yours, ours? Sharing data on human genetic variation
133	TIER2/OSIRIS	Narang (2023)	Data Sharing for Pediatric Clinical Trials Funded by the US National Institutes of Health
134	TIER2/OSIRIS	Naudet (2018)	Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in the BMJ and PLOS Medicine
135	TIER2/OSIRIS	Naudet (2018)	Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in the BMJ and PLOS Medicine
136	TIER2/OSIRIS	Nguyen (2022)	Changing patterns in reporting and sharing of review data in systematic reviews with meta-analysis of the effects of interventions: Cross sectional meta-research study

137	TIER2/OSIRIS	Nguyen (2022)	Changing patterns in reporting and sharing of review data in systematic reviews with meta-analysis of the effects of interventions: Cross sectional meta-research study
138	TIER2/OSIRIS	Nuijten (2017)	Journal data sharing policies and statistical reporting inconsistencies in psychology
139	TIER2/OSIRIS	Nuijten (2017)	Journal data sharing policies and statistical reporting inconsistencies in psychology
140	TIER2/OSIRIS	Nuijten (2017)	Journal data sharing policies and statistical reporting inconsistencies in psychology
141	TIER2/OSIRIS	Obels (2020)	Analysis of open data and computational reproducibility in registered reports in psychology
142	TIER2/OSIRIS	Palmer (2021)	Intervention reporting of clinical trials published in high-impact cardiology journals: effect of the TIDieR checklist and guide
143	TIER2/OSIRIS	Piwovar (2010)	Public sharing of research datasets: a pilot study of associations
144	TIER2/OSIRIS	Piwovar (2011)	Who shares? Who doesn't? Factors associated with openly archiving raw research data
145	TIER2/OSIRIS	Piwovar (2011)	Who shares? Who doesn't? Factors associated with openly archiving raw research data
146	TIER2/OSIRIS	Riedel (2022)	Including Data Management in Research Culture Increases the Reproducibility of Scientific Results
147	TIER2/OSIRIS	Rowhani-Farid (2016)	Has open data arrived at the British Medical Journal (BMJ)? An observational study
148	TIER2/OSIRIS	Rowhani-Farid (2018)	Badges for sharing data and code at Biostatistics: an observational study
149	TIER2/OSIRIS	Rowhani-Farid (2018)	Badges for sharing data and code at Biostatistics: an observational study
150	TIER2/OSIRIS	Rowhani-Farid (2020)	Did awarding badges increase data sharing in BMJ Open? A randomized controlled trial
151	TIER2/OSIRIS	Ruamviboonsuk (2023)	Data sharing implementation in top 10 ophthalmology journals in 2021
152	TIER2/OSIRIS	Salameh (2019)	Completeness of Reporting of Systematic Reviews of Diagnostic Test Accuracy Based on the PRISMA-DTA Reporting Guideline
153	TIER2/OSIRIS	Savage (2009)	Empirical study of data sharing by authors publishing in PLoS journals
154	TIER2/OSIRIS	Shamliyan (2014)	Availability of results from clinical research: failing policy efforts
155	TIER2/OSIRIS	Shen (2023)	The unknown lipids project: harmonized methods improve compound identification and data reproducibility in an inter-laboratory untargeted lipidomics study
156	TIER2/OSIRIS	Siebert (2020)	Data-sharing recommendations in biomedical journals and randomised controlled trials: an audit of journals following the ICMJE recommendations
157	TIER2/OSIRIS	Sims (2016)	Shoulder Arthroplasty Trials Are Infrequently Registered: A Systematic Review of Trials
158	TIER2/OSIRIS	Sims (2018)	Trial registration and adherence to reporting guidelines in cardiovascular journals
159	TIER2/OSIRIS	Sinaci (2023)	A Data Transformation Methodology to Create Findable, Accessible, Interoperable, and Reusable Health Data: Software Design, Development, and Evaluation Study
160	TIER2/OSIRIS	Soderberg (2021)	Initial evidence of research quality of registered reports compared with the standard publishing model
161	TIER2/OSIRIS	Speich (2023)	Reminding Peer Reviewers of Reporting Guideline Items to Improve Completeness in Published Articles: Primary Results of 2 Randomized Trials
162	TIER2/OSIRIS	Struthers (2021)	GoodReports: developing a website to help health researchers find and use reporting guidelines
163	TIER2/OSIRIS	Thelwall (2017)	Do journal data sharing mandates work? Life sciences evidence from Dryad
164	TIER2/OSIRIS	Toth (2021)	Study Preregistration: An Evaluation of a Method for Transparent Reporting
165	TIER2/OSIRIS	Toth (2021)	Study Preregistration: An Evaluation of a Method for Transparent Reporting

166	TIER2/OSIRIS	Toth (2021)	Study Preregistration: An Evaluation of a Method for Transparent Reporting
167	TIER2/OSIRIS	Trisovic (2022)	A large-scale study on research code quality and execution
168	TIER2/OSIRIS	Trisovic (2022)	A large-scale study on research code quality and execution
169	TIER2/OSIRIS	van der Braak (2022)	The score after 10 years of registration of systematic review protocols
170	TIER2/OSIRIS	Vassar (2020)	Assessment of the completeness of intervention reporting of randomized clinical trials for alcohol use disorders: Effect of the TIDieR checklist and guide
171	TIER2/OSIRIS	Veroniki (2021)	Do reporting guidelines have an impact? Empirical assessment of changes in reporting before and after the PRISMA extension statement for network meta-analysis
172	TIER2/OSIRIS	Walsh (2020)	Adherence to COREQ Reporting Guidelines for Qualitative Research: A Scientometric Study in Nursing Social Science
173	TIER2/OSIRIS	Witwer (2013)	Data submission and quality in microarray-based MicroRNA profiling

Authors: Panagiotis Kavouras, Pooja Aryal, Ilse de Jong, Joeri Tjldink

Reviewer: Vivian Nchanchou Mbanya, Rosemarie de La Cruz Bernabe