

TRUSTparenthood

Deliverable 3.2

Draft version of practices

Authors: Joeri Tijding, Mariska Leeflang, Pooja Aryal, Ilse de Jong, Vivian Nchanchou Mbanya, Panagiotis Kavouras

Editor: Panagiotis Kavouras

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

Project acronym: TRUSTparenthood

Grant Agreement no.: 101188172

Lead contractor for this deliverable: Stichting Amsterdam UMC (AUMc)

Deliverable factsheet:

Project Number:	134601135783
Project Acronym:	TRUSTparenCy
Project Title:	Increasing reproducibility through the co-creation of interventions that support a transparent and trustworthy research ecosystem
Title of Deliverable:	Draft version of practices
Work Package:	WP3
Authors:	Joeri Tijding, Mariska Leeftang, Pooja Aryal, Ilse de Jong, Vivian Nchanchou Mbanya, Panagiotis Kavouras
Editor:	Panagiotis Kavouras
Reviewers:	Signe Mesinska, Daniele Fanelli

ABSTRACT:	With this report we aim to provide Research Performing Organisations (RPOs), research funders, publishers, and learned societies with a transparent, defensible, and usable basis for action on reproducibility. It sets out how a large set of reproducibility interventions was taxonomized, assessed for the strength of its underlying evidence, and organized into an openly available and interactive online repository of reproducibility interventions (named in this report as "Toolbox"). This report explicates one of the three elements that comprise the institutional practices for promoting reproducibility, with the other two being the training materials and the Reproducibility Promotion Plans (RPPs). Reproducibility interventions were compiled from two independent scoping reviews (TIER2/OSIRIS and TRUSTparenCy), were described with a harmonized terminology, yielding a combined list of 172 interventions. A three-round Delphi study produced a consensus taxonomy of five categories spanning the empirical cycle, comprising 18 subcategories, with an additional "Infrastructure" category for foundational, pre-research-cycle interventions reaching 81% consensus. Each intervention type was then independently double-assessed by two teams for evidence quality and acceptability, drawing on GRADE, SOPs4RI, and iRISE outputs. The central finding is that the evidence base for reproducibility interventions remains rather weak. Most intervention types were supported by too few studies, and even the better-studied ones (i.e. reporting checklists, journal data-sharing and reporting policies, preregistration) showed only weak or mixed effects in increasing reproducibility, and scarce experimental designs, high heterogeneity. Acceptability varied by cost, effort, and who bears the burden of implementation.
Keyword List:	Reproducibility, Toolbox, Delphi, taxonomy, interventions

Consortium

	Role	Name	Short Name	Country
1.	COO	Universitetet i Oslo	UIO	NO
2.	BEN	Aarhus Universitet	AU	DK
3.	BEN	Science of Science Limited	SoS	UK
4.	BEN	Stichting Amsterdam UMC	AUMC	NL
5.	BEN	Latvijas Universitāte	UOL	LV
6.	BEN	Ethnicon Metsovion Polytechnion	NTUA	EL
7.	BEN	Tieteellisten Seurain Valtuuskunnasta	TENK	FI
8.	BEN	Aristotelio Panepistimio Thessalonikis	AUTH	EL
9.	BEN	Universidade Federal de Santa Catarina	UFSC	BR
10.	BEN	National University Incorporated	NU	PH
11.	BEN	Kwame Nkrumah University of Science and Technology Kumasi	KNUST	GH
12.	BEN	Lviv Polytechnic National University	LPNU	UA
13.	AP	Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek TNO	OSNL	NL

Revision History

Version	Date	Revised by	Reason
0.1	18.06.2026	Joeri Tjldink	Draft plan outlining scope and methodology
0.2	05.08.2026	Joeri Tjldink, Panagiotis Kavouras	First draft of the report for initial comments
0.3	14.08.2026	Signe Mesinska, Daniele Fanelli	Comments on the first draft of the report
0.4	27.08.2026	Joeri Tjldink, Mariska Leeftang	Second draft of the report, for comments
0.5	30.08.2026	Signe Mesinska	Comments on the second draft of the report
0.6	31.08.2026	Joeri Tjldink	Third draft of the report for the coordinator
0.7	03.09.2026	Panagiotis Kavouras	Comments on the third draft of the report
0.8	06.09.2026	Joeri Tjldink, Pooja Aryal	Fourth draft of the report
0.9	08.09.2026	Daniele Fanelli	Comments on the fourth draft of the report
1.0	09.09.2026	Mariska Leeftang, Ilse de Jong	Fifth draft of the report

1.1	11.09.2026	Panagiotis Kavouras	Final edits and compilation of the final version of the report
------------	------------	---------------------	--

Table of contents

1. Introduction	5
2. Harmonising the descriptions of interventions	6
3. Results of the Delphi study	7
4. How interventions are placed in the taxonomy	10
5. Assessment of interventions.....	11
5.1 Information on assessment.....	12
5.2 Output of the assessment.....	15
6. Structure and functionalities of the Toolbox.....	19
6.1 Content of the Toolbox	19
6.2 Structure of the Toolbox	19
6.3 Functionalities (back-end) of the Toolbox	21
6.4 Front-end of the Toolbox.....	21
7 Annexes	24
A. Combined list of interventions from TIER2/OSIRIS and TRUSTparenity projects with the description of each intervention.....	24
B. Score sheet and explanations	32
C. Assessment sheet	35

1. Introduction

TRUSTparency aspires to encourage researchers and other stakeholders to adopt reproducibility interventions and institutional practices that render them applicable to the day-to-day research practice to the extent that they have been provenly developed, deployed, assessed and corrected transparently and collaboratively. The mission of the project follows from this: to apply a maximally informed, context-sensitive, collaborative, democratic and transparent approach to facilitating the development of interventions and institutional practices, like the Reproducibility Promotion Plans (RPPs), that increase the reproducibility and replicability of research results.

This report presents in which way the set of reproducibility interventions have been taxonomised according to an evidence based and concise set of descriptors, assessed for their quality, and streamlined in a user-friendly way via a openly available online platform (the Reproducibility Toolbox) that can help institutions foster replicability and reproducibility and are one of the pillars of the Reproducibility Promotion Plans (RPPs) that have been developed to help organisations strengthen reproducibility by planning, implementing, assessing, and improving concrete interventions in everyday research practice. This Toolbox is one of the three elements of the project's institutional practices for reproducibility, together with the training material and the RPPs that will help Research Performing Organisations (RPO), Research Funding Organisations (RFO), or Publishers (incl. journals) develop institutional reproducibility procedures that are relevant to their specific disciplinary context, responsibilities, and communities.

The taxonomized set of reproducibility interventions, the quality assessment framework and its application to the reproducibility interventions collected by TRUSTparency and TIER2 and OSIRIS projects, the structure and functionalities of the reproducibility Toolbox, as well as the development process behind them, are presented in this report, which continues as follows:

1. We have conducted a Delphi Study (D1.1), to taxonomize the different steps of reproducibility, to organize the practices that are collected in the scoping review
2. We have assessed the quality of all the intervention types (those found by TRUSTparency and those found by the OSIRIS/TIER2 scoping reviews) to give institutions a better understanding how these interventions can be used when writing an RPP.
3. We end this deliverable with an overview of the structure and functionality of the reproducibility Toolbox.

The collection builds on the state-of-the-art knowledge base created in WP1 and reported in D1.1. Existing interventions were collected through a scoping review conducted according to PRISMA guidelines, covering peer-reviewed and grey literature, the outputs of related EU-funded projects including TIER2, OSIRIS and iRISE, institutional and transnational guidelines from RPOs, RFOs, publishers and learned societies, and documents from other initiatives. Each collected intervention was assessed on strength of evidence, feasibility,

implementability, usability and methodological rigour, following the methodology used for the quality assessment of the intervention.

The goal of this work is to provide a transparent, defensible and usable basis for institutional action on reproducibility, and to put that basis in a form the pilot institutions can immediately work with. Concretely, it aims to:

1. present the interventions in a toolbox, identified through the review, organised in a taxonomy that makes them comparable across stages of the research cycle;
2. report the 'final outcome of the Delphi study
3. make the assessment of each intervention type explicit, including the strength of the underlying evidence, its feasibility and its implementability;
4. describe the structure and functionality of the draft toolbox, and how institutions will navigate it to arrive at interventions that fit their own context;
5. provide the nine pilot institutions with a working version they can apply and criticise, so that the final version delivered in WP4 is grounded in implementation experience across different types of institutions, countries and disciplines.

2. Harmonising the descriptions of interventions

The descriptions of all interventions gathered from the two independent scoping reviews of TIER2/OSIRIS and TRUSTparency (that was conducted in the context of Task 1.1 and reported in D1.1, see: <https://trustparency-project.eu/deliverables/>) were harmonized against those utilised in the TIER2/OSIRIS scoping review. Subsequently, the combined set of reproducibility interventions were taxonomized into categories and subcategories resulting from the Delphi survey that was conducted in the context of Task 1.2 and reported in D1.1. This process of placement or taxonomizing is explicated in Section 4.

Since the scoping reviews from TIER2/OSIRIS and TRUSTparency were conducted independently, identical interventions were often described using different terminology. For example, "*journal endorses reporting checklist*" in one review and "*use of reporting checklist*" in the other, both meaning a journal-level reporting-checklist mandate. The harmonization process involved matching each TRUSTparency intervention description to its closest counterpart among the interventions identified by TIER2/OSIRIS, producing a shared vocabulary of interventions across both scoping reviews. This step contributed to a meaningful downstream classification; without a unified vocabulary, substantively identical interventions that were differently worded would be miscounted as distinct interventions, distorting the subsequent tally of evidence by intervention type. Ultimately, harmonization narrowed down and standardised how interventions were described and compared across the combined evidence base, before mapping them into the Delphi-derived taxonomy categories and sub-categories.

The combined and harmonized list of 172 interventions identified through the TIER2/OSIRIS and TRUSTparency scoping reviews can be found in Annex A.

3. Results of the Delphi study

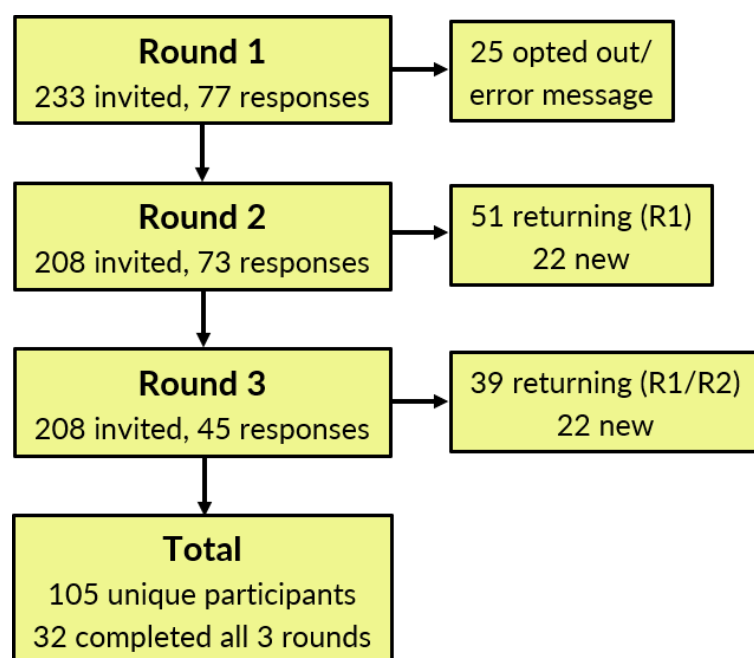


Figure 1: Participant flow across the three Delphi rounds

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.

We conducted a three-round Delphi study to create an expert-informed taxonomy of interventions for facilitating reproducible research practices at relevant stages of the research cycle. The full methodology and process of the first two rounds of the Delphi study are described in detail in D1.1. Here we report the results of the third and final round of the Delphi study and report the finalised taxonomy.

Of 233 invited panelists, 105 completed at least one round (overall response rate: 45%), and 32 participated in all three rounds. Figure 1 shows the invitation process and participation per round. **Table 1** shows the

Table 1: Characteristics of panelists who participated in at least one round

Category	N	Percentage
Gender		
Female	45	43%
Male	59	56%
Non-binary	0	0%
Prefer not to disclose	1	1%
None of the above	0	0%
Geographic region		
Africa	7	7%
Asia	0	0%
Europe	84	80%
Latin America	5	5%
North America	5	5%
Oceania	3	3%
Did not specify	1	1%

Discipline		
Health sciences	47	45%
Natural sciences	17	16%
Social sciences	32	30%
Other	9	9%
Experience with reproducibility		
Not experienced at all	1	1%
Slightly experienced	7	7%
Moderately experienced	36	34%
Very experienced	44	42%
Extremely experienced	17	16%
Role*		
Researcher	85	81%
Research support staff	13	12%
Research manager or administrator	14	13%
Journal editor or publishing professional	17	16%
Member of reproducibility network	55	52%
Member of learned society	10	10%
Policy maker	3	3%
Other	5	5%

*Panelists could select more than one role.

Based on consensus across the three rounds, we identified five categories of reproducibility interventions across the empirical cycle, comprising 18 subcategories in total. Figure 2 shows the final taxonomy. **Table 2** provides full descriptions of all subcategories as formulated through the three Delphi rounds.

In this third and final round the first suggestion was to add an "Infrastructure" category for foundational, pre-research-cycle interventions implemented by funders, research performing organisations and publishers (such as data management plan mandates), rather than by researchers during the research cycle. This suggestion reached 81% consensus; panelists strongly supported the idea that adequate infrastructure is essential for reproducible practice, though several objected that the word "infrastructure" is easily confused with technical facilities like high-performance computing or laboratory freezers.

The second suggestion was to rename the "Write the report" stage to "Disseminate" and fold "Post-publication" and "Peer-review" into it as subcategories, which reached 74% consensus. Here the main objection was to the term "disseminate" (interpreted by some as only beginning after publication) and to combining conceptually distinct activities, since peer review is largely pre-publication and merging reporting, validation and diffusion risks blurring them; the steering committee acknowledged this but argued the subcategories provide sufficient granularity.

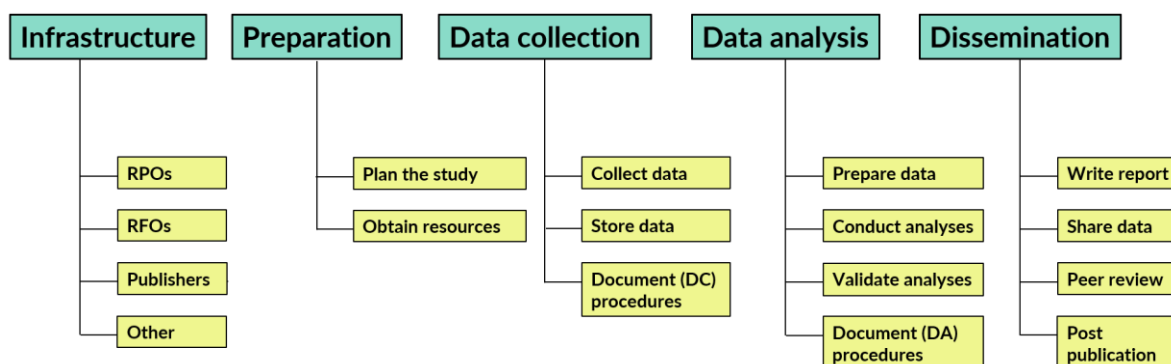


Figure 2: Final taxonomy of reproducibility interventions. DC stands for “Data Collection” and DA stands for “Data analyses”.

Table 2: Descriptions of subcategories after three Delphi rounds.

Infrastructure	
<i>Research Performing Organizations (RPOs)</i>	Organizations that perform research and provide resources, policies, and support for their researchers
<i>Research Funding Organizations (RFOs)</i>	Organizations that fund researchers and set requirements or incentives for reproducibility
<i>Publishers (incl. journals)</i>	Organizations that disseminate research and set standards for transparency and reproducibility in publication
<i>Other</i>	intended to accommodate interventions that clearly belong within the <i>Infrastructure</i> category, but do not align with any of the existing subcategories
Preparation	
<i>Formulate the research question/hypothesis</i>	The process of defining what the study aims to investigate
<i>Plan the study</i>	Creating a plan for how the study will be conducted
<i>Obtain resources</i>	Securing the resources necessary to execute the study
Data collection (DC)	
<i>Collect data</i>	Executing the data-gathering process
<i>Store data</i>	Saving and organizing the collected data
<i>Document data collection procedures</i>	Recording details of how the data are collected
Data analysis (DA)	
<i>Prepare data</i>	Organizing and cleaning the data to prepare for analysis
<i>Conduct analyses</i>	Conducting quantitative and/or qualitative analysis
<i>Validate analyses</i>	Assessing the accuracy and reliability of findings
<i>Document DA procedures</i>	Recording details of how the data are analyzed
Dissemination	
<i>Write report</i>	Report the findings of the study
<i>Share data</i>	Making research data accessible to others
<i>Peer-review</i>	The evaluation of research by others
<i>Post-publication</i>	Contributing to the ongoing use, evaluation, and discussion of research after publication

4. How interventions are placed in the taxonomy

All studies identified through the TIER2/OSIRIS and the TRUSTparenCy scoping reviews were taxonomised into the categories and subcategories informed by the Delphi consensus survey. This process was executed by Ilse de Jong (AUMC) and Pooja Aryal (UiO). Studies for which categorization was not agreed upon were categorised by Joeri Tjink (AUMC) and Panagiotis Kavouras (UiO).

Within each subcategory, studies were further grouped into intervention types, meaning that studies examining similar interventions (e.g. reporting guidelines, data availability statements) were labeled under the same type. This intervention type level formed the basis for the subsequent assessment procedure. As some studies examined multiple interventions, a single study could contribute evidence to multiple intervention types. Intervention type categorization was carried out by Ilse de Jong (AUMC), and checked by Pooja Aryal (UiO), Panagiotis Kavouras (UiO), Mariska Leeftang (AUMC), and Joeri Tjink (AUMC). Disagreements were resolved through group discussion. **Table 3** shows the final placement of interventions into different types and the number of published studies identified for each of these types.

The identification of intervention types within subcategories was based solely on the interventions identified in the two scoping reviews. The current set of intervention types should therefore be seen as a reflection of the evidence base at the time this report was compiled, rather than the full spectrum of possible interventions.

Table 3: Categorization of reproducibility interventions.

Category	Subcategory	Intervention type	N (studies)
Infrastructure	Publishers	Journal policy on completeness of reporting	18
		Journal policy on data sharing	30
		Journal policy on registration	7
		Journal awards open science badges	8
		Results-free peer-review	1
		Outcome-agnostic acceptance by journal	1
	RFOs	Funder policy on data management	3
	RPOs	Research institution provides support services	2
		Research institution provides training	3
		Research institution provides role models	1
	Other	Government policy on registration	2
		Data repository	3
Preparation	Plan the study	Preregistration	6
		Optimization of study design	5
		Statistical analysis plan	1
		Standardization of experimental protocol	2
	Obtain resources	Researcher trained on study protocol	3
Data collection	Collect data	Researcher uses standard method for data acquisition	11
		Multi-site testing	1

	Store data	Standardized file system structure	3
		Data storage solution	4
		Direct data deposition	1
	Document data collection procedures	Researcher uses standardized documentation framework	1
		Software tool that documents data retrieval procedures	2
		Real-time procedure documentation	1
Data analysis	Prepare data	Standardized data preparation procedure	5
		Data quality control	1
	Conduct analyses	Researcher uses common method for data analysis	19
		Tool to standardize data analysis procedures	2
	Validate analysis	Multiverse analysis	1
		Authors use effect-size filter	3
	Document data analysis procedures	Researcher uses standardized documentation framework	2
		Open workflow platform	3
		Researchers use tool to document analysis procedures	4
Dissemination	Write the report	Use of reporting checklist	15
	Share data	Researchers share data	4
		Data management tool	5
	Peer-review	Peer-reviewer checks reporting	4
	Post-publication	Post-publication data requests	2
		Discrepancy review	1
		Replication	3

5. Assessment of interventions

To decide on the implementation of practices (interventions) to promote transparency and reproducibility of science, it is important to know if these practices indeed improve reproducibility. In other words, is there evidence that these interventions are effective? The answer to this question should come from empirical research implemented by the authors of the collected peer reviewed publications that studies the interventions.

The “gold standard” study design to prove that an intervention is effective, is – according to biomedical methodology – a (randomized) controlled experiment. This is arbitrary, but given that there is no other gold standard, we used this). Alternatively, other comparative designs may be used. In any case, the effect of doing something should be compared to what happens if that ‘something’ is not done. Uncontrolled studies do not provide a comparison between the situation with and without the practice and thus should be treated with caution. Opinions cannot be considered as evidence, although they may be important for decision making. In that case, one should be transparent about the decision being mainly opinion-based. Observational comparative designs may suffer from confounding bias. For example, if the effect of introducing data stewards

is investigated between institutes with and without data stewards, then the institutes with data stewards may already have a more stricter reproducibility infrastructure, so that the apparent effect of the data stewards is actually caused by the total infrastructure.

To provide a scientific basis for the effectiveness of many interventions that may improve the reproducibility of research, TIER2, OSIRIS, and iRISE EU-funded projects have collected all available evidence on such interventions in scoping reviews. However, the output of these scoping reviews was not formally assessed for its quality. Here, we present a structure to assess the validity and value of the interventions collected by the TIER2/OSIRIS and TRUSTparenacy scoping reviews, based on previous work (SOPs4RI, iRISE, and GRADE¹). Below, the assessment criteria are presented; the summary of the actual assessment is done in **Table 4** (see below).

5.1 Information on assessment

The assessment of the body of evidence for each intervention was done by two persons, independent from each other, from 2 different teams (AUMC and UiO).

The first part of the assessment sheet () consists of basic details of the interventions that will be assessed. Among the team, there was some discussion about what an intervention constitutes: is it an interventional study, is it a very specific practice, or a class of research practices? We defined an intervention as a practice that can be done to improve reproducibility. This practice may be defined very specifically (e.g. preregistration in a freely accessible repository without quality checks) or more broadly (e.g. any form of providing details about the study to a broader audience before starting data collection).

Every intervention is evaluated in one or more empirical studies. The number of studies evaluating this intervention is recorded, as well as the outcome that these studies used. Interventions evaluated in only 1 study were considered to have too low certainty of the evidence to make firm recommendations about them. The ultimate outcome is reproducibility of research, but some studies may have used proxy outcomes such as data sharing or transparent reporting.

Basic details also include whether a systematic review or meta-analysis on the intervention was found, and whether quantitative results are available. If a (recent = <5 years) systematic review or meta-analysis was available, then only one report (the review) can be sufficient. A meta-analysis may also provide quantitative results.

Finally, the conclusions are stated. If all relevant studies in one type of intervention category point in the same direction, then we can use one answer for all studies. If the studies differ, then please explain if a “majority vote” was chosen and whether another choice was made. Next, the studies were assessed for their quality, according to the items in **Table 4**.

¹ Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011 Apr;64(4):380-2. doi: 10.1016/j.jclinepi.2010.09.011.

Table 4: Assessment criteria for evidence judgement.

Item	Description / Explanation	Judgment
Comparative experimental design	Indicate whether the intervention has been assessed in at least one study with an experimental design in which the effect of the intervention was compared against the effect of another or no intervention	☐ Yes – 1 study ☐ Yes: more than 1 study ☐ No
Comparative studies	Indicate whether the majority of the studies compared the effects of an intervention against the situation where no or another intervention was used.	☐ Yes ☐ No, but some studies did ☐ No, no comparative studies
Type of comparative study		☐ Independent before-after study ☐ Parallel arms at the same time ☐ Before-after study with the same subjects before and after the intervention ☐ N/A
Prospective study design	Indicate whether most of the studies planned their assessment before the moment of the baseline measurement. For example, a study that looked back at an intervention that was implemented two years ago, does not use a prospective design	☐ Yes ☐ No, but some studies did ☐ No, no prospective studies
Imprecision – sample sizes	How many studies address this intervention?	☐ <5 ☐ 5-10 ☐ >10
Imprecision – sample sizes	What is the median (range) sample size of the included studies? <i>This can be difficult when the studies use a different unit of analysis (e.g. authors, or studies, or analyses, or journals, etc)</i>	Numbers: Median (range)
Imprecision – quantitative result	If a quantitative result was presented, how precise was this? Please describe the estimate plus confidence interval (or other measure of uncertainty)	☐ Precise ☐ Unclear ☐ Imprecise ☐ N/A

Inconsistency	Indicate whether the results of all studies point into the same direction?	☐ Yes ☐ No ☐ N/A
Variation (heterogeneity)	To what extent is the body of evidence diverse and heterogeneous? Are there concerns that the body of evidence is not applicable to our situation?	☐ Yes ☐ No ☐ Unclear

After this assessment, the assessors were asked to provide an overall judgment about the quality of the included evidence about the intervention. They are invited to provide strong points and weak points of the evidence and summarize their findings in their own words.

Besides the strength or quality of the scientific evidence, there are other considerations that play a role in decisions about whether an intervention should be implemented or not. These criteria may relate to (financial) costs, effort and time needed, whether people need to be trained, etc. and they were informed by the results of Task 1.4 that have been reported in D1.3. These criteria are much more subjective to assess, so our assessments can only be quite superficial. In **Table 5** these other assessment criteria are listed. Assessors accompanied their assessment with a justification as to why this assessment was made. These justifications could be based on individual experiences, on general knowledge, or on information presented in the relevant studies.

Table 5: Assessment criteria for acceptability judgement.

Item	Description / Explanation	Judgment
Financial costs	Some studies mention this explicitly, but most will not. Low financial costs means that minimal funds are required to implement the intervention; high costs may relate to institution-wide new infrastructures, or for example the employment of multiple new staff members. Open access may have medium or high costs, depending on the number of publications, the fees and whether someone is based in HIC or LMIC.	☐ Low ☐ Medium ☐ High
Effort and time investment	Low: the intervention costs a few hours per project extra (e.g. open access publication); Medium: the intervention requires more than a few hours (e.g. preregistration in a separate repository; or data sharing when	☐ Low ☐ Medium ☐ High

	the data have been curated along the way); High: data curation on institutional levels; open and transparent workflows.	
Required knowledge	Is training needed before someone can use / do the intervention??	☐ Yes ☐ No
Set-up and infrastructure	Specify what kind of infrastructure may be needed; provide arguments as to why you choose any of the tick boxes.	☐ Low ☐ Medium ☐ High
Potential harms or disadvantages	Please provide any potential harms or disadvantages, separate from (financial) costs	

After this assessment the assessors were asked to provide an overall judgment about the feasibility and acceptability of the intervention. They are invited to provide strong points and weak points of the intervention, as well as to summarize their opinion in their own words. Finally, both the quality of the evidence and the feasibility will be summarized in a short account of the intervention.

Each intervention was assessed by two persons independently of each other. Afterwards, the two assessments were compared by Ilse de Jong or Mariska Leeftang and disagreements were solved by either of them. Annex B presents the instructions that were given to the assessors and Annex C presents the excel sheet, where the assessment of all involved researchers was compiled by Mariska Leeftang. We did not quantify the level of agreement between assessments or discuss the assessments within the team, due to time constraints. Current assessments are therefore preliminary. In general, most disagreements were between how the term “intervention” was interpreted (specified for the included studies, or in more generic terms); how the term “outcome” was interpreted (idem); and for justifications about the acceptability-items. These disagreements will be discussed in a later stage vis-à-vis the results of pilot testing that will inform the final list of interventions that will be reported in D4.2: Final collection of practices.

5.2 Output of the assessment

The assessment process led to a comprehensive list of assessed interventions, according to the stage of research, the intervention type and the specific intervention; and each intervention is accompanied by a row of assessment details and the overall judgment. See below (**Table 6**) for an overview of all intervention types and their average judgement scores on evidence and acceptability. This score will eventually inform the decision on which intervention to choose of the end users of the toolbox.

In the coming months of the TRUSTparenity project, we will test instances of some of these intervention types in pilot institutions and their feedback on the tools and interventions will help shape the description of the chosen intervention types and their implementability. **Table 6** thus presents first draft.

Table 6: Overview of all intervention types, the number of studies found and the judgment of the intervention types by both evidence and acceptability.

Intervention type	N	Evidence judgment [#]	Acceptability judgment ^{\$}
Journal policy on completeness of reporting: journal endorses or requires use of reporting checklists	18	Positive but weak effect; few (n=3) experimental designs; high heterogeneity. Low certainty	Burden lies with publishers; commonly used; high acceptability
Journal policy on data sharing: journal asks or demands sharing of data	30	Half positive effect; half no effect; no experimental designs; some heterogeneity.	Costs and effort depend on level of mandating; research may remain reluctant to share
Journal policy on registration: journal only publishes clinical trial when this was registered beforehand	7	Positive effect; no experimental designs; high heterogeneity. Low certainty.	Low costs and efforts; researchers may be reluctant
Journal awards open science badges: author team receives badge when they make their data, code or materials open and publicly accessible.	8	Positive effect; no experimental designs; high heterogeneity. Low certainty.	Low costs and efforts; effect of badges on reproducibility seems limited.
Results-free peer-review	1	Only 1 study; very low certainty of any effect.	Costs and efforts from the peer reviewers' side are low; for the journal it changes decision making
Outcome-agnostic acceptance by journal	1	Only 1 study; very low certainty of any effect.	Costs and efforts from the peer reviewers' side are low; for the journal it changes decision making
Funder policy on data management	3	Studies investigated NIH policy and found no effect. Low certainty.	Costs and effort depend on level of mandating; researchers may be reluctant
Research institution provides support services	2	Two very different non-comparative studies show a positive effect. Low certainty.	Requires infrastructure, staff and commitment. Potentially high costs.
Research institution provides training	3	Mixed effects; 2 comparative studies; heterogeneous types of training. Low certainty.	Acceptability depends on type of training; incentives may still be needed.
Research institution provides role models	1	Only 1 study; very low certainty of any effect.	Acceptable
Government policy on registration	2	Unclear effect in two very different studies. Before-after design. Low certainty.	Requires government to take action; no longer in the hands of the scientific community
Data repository	3	Three non-comparative studies mention a positive effect. Very low certainty.	Data repositories are already accepted measures to share data

Preregistration: researcher publishes a protocol of their study (before the results are known)	6	Weakly positive effect; no experimental designs; high heterogeneity. Low certainty.	Relatively cheap and effortless; training of referees may increase costs
Optimization of study design: A wide varying range of interventions: (1) standardization / adjustment of effect sizes; (2) best lab practices; (3) Python toolbox.	5	Positive or unclear effect; 1 experimental design; very variable. Too heterogeneous to make a statement.	Too heterogeneous to make a statement.
Statistical analysis plan	1	Only 1 study; very low certainty of any effect.	Acceptable, especially when not too much technical knowledge required.
Standardization of experimental protocol	2	Two articles of which one was a (descriptive) systematic review and the other an observational comparative study. The latter showed a positive effect. Low certainty.	Development of shared standards takes time, and may not be applicable in all contexts
Researcher trained on study protocol	3	Two studies introduced structured workflow next to training. Positive effect; low certainty.	Acceptability depends on type of training; incentives may still be needed.
Researcher uses standard method for data acquisition: standardized physical protocols, harmonized experimental procedures, and quality control frameworks	11	Positive effect; 1/11 experimental designs; high heterogeneity. Low certainty.	Diverse interventions, so varying acceptability.
Multi-site testing	1	Only 1 study; very low certainty of any effect.	We think this is not really an intervention
Standardized file system structure	3	Non-comparative studies; positive effect. Low certainty.	Acceptability may depend on type of structures and level of commitment
Data storage solution	3	Three non-comparative studies showing a positive effect. Low certainty	Acceptable, but may require funding and training
Direct data deposition	1	Only 1 study; unclear effect; very low certainty of any effect.	Based on one case study, it was difficult to judge acceptability.
Researcher uses standardized documentation framework	1	Only 1 study; very low certainty of any effect.	Overlaps with other standardizations
Software tool that documents data retrieval procedures	2	To very different studies show positive effect	Too heterogeneous to assess acceptability.
Real-time procedure documentation	1	Only 1 study; very low certainty of any effect.	Based on one case study, it was difficult to judge acceptability.
Standardized data preparation procedure: A wide varying range of interventions: (1) harmonisation of data; (2) automated preparation or reporting of data; (3) standardization -omics data and analyses	5	Low certainty; although it seems logical that standardized methods improve reproducibility, these studies do not prove that	Diverse interventions, so varying acceptability.

Data quality control	1	Only 1 study; very low certainty of any effect.	Depends a lot on exact control measurements and infrastructure
Researcher uses common method for data analysis: Coordinating research centres and software developers mandating shared standardized analytical algorithms to process data	19	Low certainty; although it seems logical that standardized methods improve reproducibility, these studies do not prove that. No experimental designs.	Depends on level of standardization and mandates; it asks from all collaborators to adhere to a fixed structure.
Tool to standardize data analysis procedures	2	Comparative before-after studies; positive effect; low certainty.	Unclear
Multiverse analysis	1	Only 1 study; very low certainty of any effect.	Is this really an intervention?
Authors use effect-size filter	3	Positive effects, but not based on comparative analyses. Low certainty.	Is this really an intervention; unclear acceptability.
Researcher uses standardized documentation framework	2	Positive effect, based on anecdotal evidence. Very low certainty.	Acceptable, as these frameworks rely on codebooks and existing repositories. But may require commitment
Open workflow platform	3	Positive effect in three very different (non-comparative) studies on very different interventions. Low certainty.	Using cloud computing can lead to dependence on third-party providers.
Researchers use tool to document analysis procedures	4	Positive effect in three very different (non-comparative) studies on very different interventions. Low certainty.	Low-cost, low-to-moderate-effort tools
Use of reporting checklist: use of a reporting checklist or writing aid by individual researchers	15	Overall effective, but weak effects; 3 (quasi) experimental designs; relatively homogeneous. Moderate certainty.	Passive guideline existence alone is cheap and easy, but also less effective than active or tool-supported interventions
Researchers share data	4	Retrospective non-comparative studies showing positive effect. Low certainty.	Acceptability depends on privacy issues and effort needed to make shared data really reusable.
Data management tool: variable tools that accommodate or support data analysis, data storage, FAIR-ization, etc	5	Positive effect; no experimental designs; high heterogeneity. Low certainty.	Too heterogeneous to make a statement.
Peer-reviewer checks reporting	4	Positive effect; three randomized experiments. Moderate-high certainty that this may have an effect.	Especially a dedicated peer reviewer may cost time and effort; but overall it is feasible
Post-publication data requests	2	Comparative studies, but effect is unclear. This is a post-hoc intervention.	Low costs, low efforts; but takes action after publication

Discrepancy review	1	Positive effect in only 1 case-study. Very low certainty.	May cause discussions about study afterwards.
Replication	3	Unclear effect of a post-hoc intervention; non-comparative studies. Low certainty.	This does not improve the quality of the research prospectively.

Evidence judgment is a judgment based on the quality of the studies that reported on this intervention. The question was “How certain are we that this intervention improves reproducibility?”

\$ Acceptability judgment is based on a subjective indication of the costs, effort and required knowledge and infrastructure of the interventions. These may vary between settings and countries.

6. Structure and functionalities of the Toolbox

The TRUSTparency Toolbox of reproducibility interventions (the “Toolbox” for the rest of the chapter) has been designed to be a structured, online repository of the taxonomised and assessed set of more than 170 reproducibility interventions, openly available through the project’s website front page.

6.1 Content of the Toolbox

The content of the Toolbox, i.e. the peer-reviewed publications of specific reproducibility interventions, has been compiled by two scoping reviews: the one jointly conducted by OSIRIS and TIER2 projects and TRUSTparency’s scoping review (reported in D1.1) that was based on the methodology (like inclusion/exclusion criteria and most of the extraction data) of TIER2/OSIRIS scoping review.

6.2 Structure of the Toolbox

The structure of the Toolbox, i.e. the way the different reproducibility interventions have been organized in **Groups** and **Subgroups**, has been based on the results of the Delphi survey (reported in D1.1). As a result, the specific structure does not represent the views of the TRUSTparency consortium partners, but it is based on empirical evidence. One additional organizational level has been added, the **Types** of interventions, with the aim to break down the Subgroup category. This was decided based on: (a) the fact that some Subcategories contained too many interventions; e.g. the “Publishers” Subcategory contains 65 interventions and (b) to provide less generic and more descriptive names to smaller groups of interventions that would help the end user. The names of the **Types** of interventions have been developed by Ilse de Jong, Mariska Leefmang and Joeri Tijndink and were reviewed and finalized by Pooja Aryal and Panagiotis Kavouras. **Table 3** presents the structure of the Toolbox in tabular form, while **Figure 3** depicts the structure of the Toolbox in three treemap diagrams; the area of each sector (whether it represents Category, Subcategory or Type) is analogous to the number of interventions (i.e. peer reviewed publications) it contains.

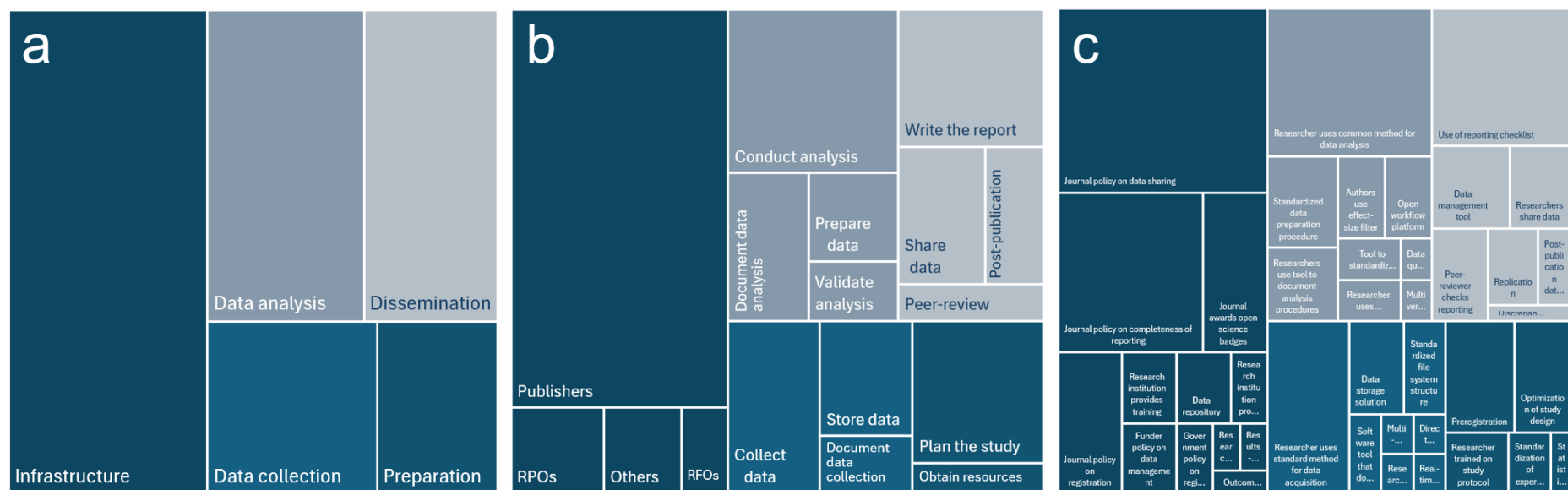


Figure 3: The structure of the Toolbox in three treemap diagrams. Structure of (a) Categories, (b) Sub-categories, and (c) Types of interventions. The area of each sector is analogous to the number of interventions it contains.

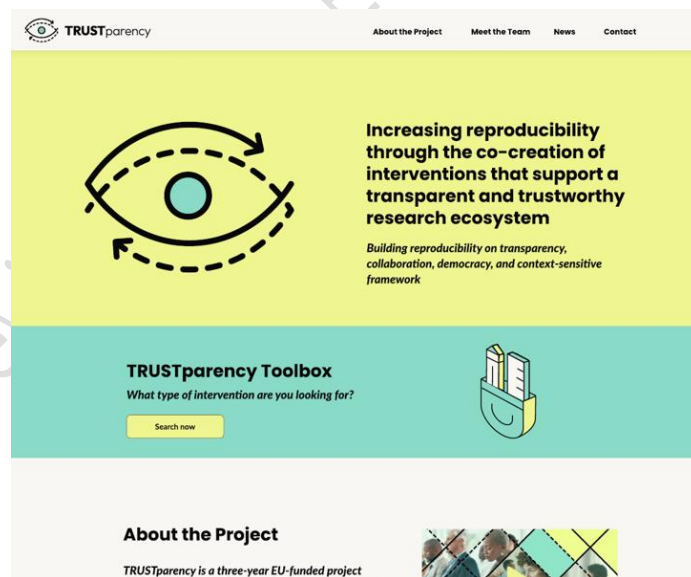


Figure 4: Placement of the Toolbox at the top of the project's front page.

6.3 Functionalities (back-end) of the Toolbox

The main aim of the Toolbox is to provide to the end user all compiled, taxonomised, and assessed reproducibility interventions in user-friendly way. This means that the interventions have to be provided in a way that the end user: (a) will be provided with a concise set of information about a reproducibility intervention, without her/him having to dig into the details of the peer-reviewed publication and (b) will be able to search the compiled interventions with an easy, interactive, and intuitive way. These two objectives defined the main **functions** of the toolbox, which subsequently defined the way the Toolbox's back end had to be developed.

For the first objective to be met, a set of descriptors for each Type intervention was compiled. These descriptors provide the following information for each Type of intervention (The final set of descriptors that will go in the toolbox is yet to be finalized, based on the results of the pilot testing):

- **Basic descriptors:** Category, Subcategory a specific Type of intervention belongs to
- **Social/epistemic descriptors:** Relevant type(s) of end-user(s), relevant research field(s)
- **Quality descriptors:** Strength of evidence, Precision, Consistency, Financial costs, necessary infrastructure (see section 5 for more information)

For the second objective to be met there was a need to develop a search functionality that the end user would not have to make too many clicks (ideally no more than three) to proceed to an ever-narrower selection of interventions (i.e. to avoid a "deep" repository); this promotes easiness of use. Also, instead of using a flow chart-like selection function, where the user would be asked to select from a specific set of alternatives (like in the case of the SOPs4RI Toolbox for Research Integrity guidelines), we used a structure of filters, where the end user will select, in one blow, all the necessary options without having to select all available options in a step-wise manner. With this strategy, the end user is also able to follow how much the number of available interventions narrows down with the use of each filter in real time; this promotes the interactive character of use. Finally, we were inspired by a well-known commercial platform, related to touristic services, on how we should present the different information hierarchies of the Toolbox. We made the following analogy between the search elements of the commercial platform and the Toolbox: Country=Category, City=Subcategory, Hotel=Type, and Hotel room=specific intervention. With this way we imitated the easiness of use of the specific platform; this promotes the intuitive character of use.

6.4 Front-end of the Toolbox

The user friendliness of the Toolbox will be based on the easiness with which the end user will be able to find the most relevant intervention to her/his interest, reflecting the architecture of the back-end. The Toolbox will be reached by a button at the top of the project's front page (**Figure 4**). Each intervention will be retrievable with the use of different filters that will appear on the left side of the Toolbox (**Figure 5**) and will have the form of a "Intervention type card" (**Figure 6**). By clicking different filters, different types of interventions will appear with the relevant interventions listed at the end of each retrieved type. Then the end user will have to click one last time to reach the peer-reviewed publication. As a result, the Toolbox has essentially two levels: the one shown at **Figure 6** that will appear after the use of filters and the actual peer-reviewed publication. As can be seen in **Figures 5 and 6**, the end user will always have a view of which filters she/he has used, thus always keeping track of the consecutive decisions she/he has taken when searching for a reproducibility intervention.

All terms that are visible in **Figures 5 and 6** will be supported with a brief explanation that will be visible when the user hovers over any of these terms with her/his cursor.

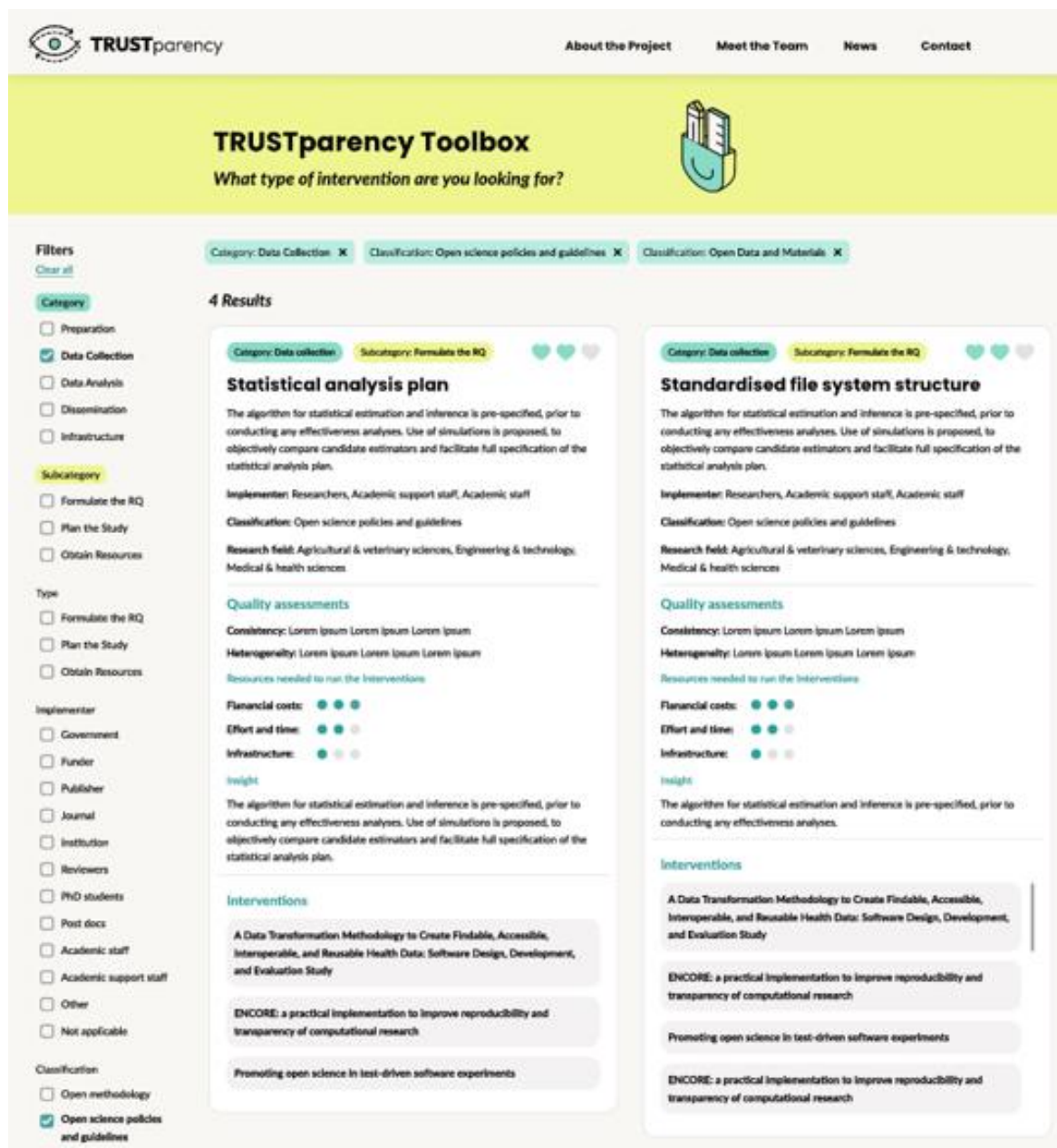


Figure 5: The Toolbox with a mock-up search with the use of filters.

Category: Data collection
Subcategory: Formulate the RQ





Statistical analysis plan

The algorithm for statistical estimation and inference is pre-specified, prior to conducting any effectiveness analyses. Use of simulations is proposed, to objectively compare candidate estimators and facilitate full specification of the statistical analysis plan.

Implementer: Researchers, Academic support staff, Academic staff

Classification: Open science policies and guidelines

Research field: Agricultural & veterinary sciences, Engineering & technology, Medical & health sciences

Quality assessments

Consistency: Lorem ipsum Lorem ipsum Lorem ipsum

Heterogeneity: Lorem ipsum Lorem ipsum Lorem ipsum

Resources needed to run the Interventions

Financial costs: ● ● ●

Effort and time: ● ● ●

Infrastructure: ● ● ●

Insight

The algorithm for statistical estimation and inference is pre-specified, prior to conducting any effectiveness analyses. Use of simulations is proposed, to objectively compare candidate estimators and facilitate full specification of the statistical analysis plan.

Interventions

A Data Transformation Methodology to Create Findable, Accessible, Interoperable, and Reusable Health Data: Software Design, Development, and Evaluation Study

ENCORE: a practical implementation to improve reproducibility and transparency of computational research

Figure 6: An “intervention Type card”.

7 Annexes

A. Combined list of interventions from TIER2/OSIRIS and TRUSTparenacy projects with the description of each intervention

No.	First author	Category	Subcategory	Intervention exact
1.	Abdelaziz (2023)	Open methodology	on data/code sharing or open science practices	Standardization of processes
2.	Aguilar Gomez (2023)	Open science tools	open workflow tools, workflow, management systems	Evaluation tool/FAIR principles
3.	Ahern (2024)	Open data and materials	data availability	Statistical tool or intervention
4.	Anido-Alonso (2023)	Open data and materials	data availability	Data/code sharing/privacy-preserving method
5.	Ashkarran (2024)	Open methodology	open protocol	Standardization guidelines
6.	Bahl (2023)	Open methodology	project workflow	Workflow tool/FAIR workflow
7.	Bontempi (2024)	Open science tools	open workflow tools, workflow, management systems	Workflow tool
8.	Bramley (2024)	Open data and materials	data availability	Editorial request for code sharing
9.	Chen (2025)	Open methodology	open protocol	Standardization guidelines
10.	Denisova (2024)	Open methodology	open protocol	Methodological guidelines/reporting guidelines
11.	DeVries (2024)	Open science tools	open workflow tools, workflow, management systems	Workflow tool/software tool
12.	Ellis (2024)	Open science policies and guidelines	policy guidelines (e.g. of funders/publishers)	Journal policy impact
13.	Elmanzalawi (2025)	Open data and materials	data availability	Software tool/workflow tool
14.	Fernholz (2024)	Open science tools	open workflow tools, workflow, management systems	Workflow tools/open software
15.	Ferreira (2023)	Open methodology	project workflow	methodological guidelines/reporting guidelines
16.	Fiedler (2023)	Open methodology	open protocol	Standardization guidelines
17.	Filatovas (2024)	Open science tools	open workflow tools, workflow, management systems	Technical/reproducibility tools
18.	Fraser (2023)	Open methodology	open protocol	Meta research/methodological tool
19.	Gaignard (2023)	Open data and materials	data availability	Data/code sharing/FAIR assessment tool
20.	Gierend (2023)	Open data and materials	data availability	Data infrastructure
21.	Gillies (2024)	Open methodology	open protocol	Harmonized protocol implementation
22.	Goetze (2024)	Open	open protocol	Workflow tools/analytical methods

		methodology		
23.	Hall (2023)	Open science tools	open workflow tools, workflow, management systems	Statistical tool
24.	Haven (2023)	Open science policies and guidelines	on data/code sharing or open science practices	Role model/supervision as intervention
25.	Hayashi (2023)	Open science tools	open workflow tools, workflow, management systems	Workflow tools/open cloud platform
26.	Hellen (2023)	Open science tools	open workflow tools, workflow, management systems	Workflow tools/software
27.	Huckvale (2023)	Open data and materials	data availability	Data/code sharing/benchmarks
28.	Jaric (2024)	Other		Experimental evaluation
29.	Kalvaityte (2024)	Open methodology	open protocol	Standardization technique
30.	Kang (2023)	Open methodology	open protocol	Statistical/study design intervention
31.	Kekecs (2023)	Open methodology	pre-registration, study registration, analysis plan	Open science intervention
32.	Kessel (2024)	Open methodology	open protocol	Standardization technique/workflow tools
33.	Kolokathis (2024)	Open science tools	open workflow tools, workflow, management systems	Workflow tools
34.	Krahmer (2023)	Open science policies and guidelines	on data/code sharing or open science practices	Behavioral/policy intervention
35.	Landau (2024)	Open methodology	open protocol	Standardization technique/workflow tools
36.	Liu (2024)	Open science policies and guidelines	reporting guidelines, reporting standards	Methodological guidance
37.	Li (2024)	Open science tools	open workflow tools, workflow, management systems	Workflow automation tool
38.	Llovera (2024)	Open methodology	open protocol	Standardization procedures/training
39.	Luna (2024)	Open methodology	open protocol	Experimental protocol/workflow tools
40.	Lusa (2024)	Open science policies and guidelines	on reporting checklists or guidance	Methodological guidance/workflow tools
41.	Mager (2023)	Open science policies and guidelines	on reporting checklists or guidance	Protocol standardization
42.	Martinc (2024)	Open data and materials	data availability	Workflow tool/open platform
43.	Martin (2023)	Open data and materials	data availability	Data infrastructure/federated analysis
44.	Miyazaki (2023)	Open science policies and guidelines	on research quality	Standardization of processes
45.	Nance (2024)	Open methodology	project workflow	statistical tool or intervention

46.	Omondi (2024)	Open science policies and guidelines	reporting guidelines, reporting standards	statistical method/workflow tool
47.	Payares-Garcia (2023)	Open science policies and guidelines	on data/code sharing or open science practices	standardization of processes
48.	Pelletier (2024)	Open science policies and guidelines	on research quality	statistical intervention
49.	Rehal (2023)	Open science policies and guidelines	on data/code sharing or open science practices	methodological intervention
50.	Rengasamy (2023)	Other		Analytical intervention
51.	Rieger (2024)	Open science policies and guidelines	on data/code sharing or open science practices	Computational reproducibility tool
52.	Rojas-Velazquez (2024)	Open science policies and guidelines	on reporting checklists or guidance	Machine learning framework
53.	Schacherer (2023)	Open science policies and guidelines	on data/code sharing or open science practices	FAIR compliant platform
54.	Schuetz (2023)	Open methodology	open protocol	Software tool
55.	Senior (2023)	Open science policies and guidelines	on research quality	Statistical simulation design
56.	Souza (2023)	Open science policies and guidelines	on reporting checklists or guidance	Standardized workflow
57.	Sprang (2024)	Open science policies and guidelines	on data/code sharing or open science practices	Data quality control
58.	Stahl (2023)	Open science policies and guidelines	reporting guidelines, reporting standards	Reporting standards and guideline implementation
59.	Tang (2024)	Open science policies and guidelines	on data/code sharing or open science practices	Data normalization method
60.	Tång-Hallbäck (2025)	Open science policies and guidelines	reporting guidelines, reporting standards	Surveillance standardization
61.	van Houdt (2023)	Open science policies and guidelines	on data/code sharing or open science practices	Open-source software
62.	Van Kampen (2024)	Open science policies and guidelines	on data/code sharing or open science practices	Standardizing computational reproducibility tool
63.	Viant (2024)	Open science policies and guidelines	on research quality	Categorization as demonstration of reproducibility
64.	Wallace (2024)	Open science policies and guidelines	reporting guidelines, reporting standards	data harmonization/standardization
65.	Wirz (2024)	Open methodology	project workflow	Data labelling
66.	Yousefirizi (2024)	Open science tools	open workflow tools,	Evaluation framework

			workflow, management systems	
67.	Zhang (2023)	Open methodology	open protocol	Protocol standardization
68.	Zhu (2023)	Open science policies and guidelines	reporting guidelines, reporting standards	Documentation standardization
69.	Agha (2016)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal endorses reporting checklist
70.	Agha (2017)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
71.	Alsheikh-Ali (2011)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines
72.	Arroyo-Araujo (2019)	Open methodology	Project workflow	standardization of processes
73.	Askarov (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data/code sharing policy
74.	Audet (2017)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE endorses reporting guideline
75.	Auspurg (2020)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data/code sharing policy
76.	Barnes (2015)	Open science tools	Other	writing tool
77.	Bergeat (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE data sharing statement
78.	Bergeat (2022)	Open data and materials	Data availability	authors shared data
79.	Blanco (2020)	Other	Other	editorial checks/peer review
80.	Blanco-Perez (2020)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal accepts regardless outcome
81.	Can (2011)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
82.	Caron (2017)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal endorses reporting checklist
83.	Caron (2017)	Open science policies and guidelines	Reporting guidelines, reporting standards	use of reporting checklist
84.	Chan (2023)	Open science tools	Dockerization, docker	statistical tool
85.	Chow (2018)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
86.	Clayson (2019)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
87.	Cobo (2011)	Other	Other	editorial checks/peer review
88.	Crüwell (2023)	Open science policies and guidelines	Rewards and incentives	badges for open data

89.	Culina (2020)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal code sharing policy
90.	Danchev (2021)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE data sharing statement
91.	Deardorff (2020)	Open educational resources	Training	training
92.	Federer (2018)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
93.	Fernandes (2019)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of methodological guidelines
94.	Findley (2016)	Open evaluations	Other	results-free review
95.	Fuller (2014)	Open science policies and guidelines	Reporting guidelines, reporting standards	use of reporting checklist
96.	Giofre (2017)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines
97.	Giofre (2017)	Open science policies and guidelines	Rewards and incentives	badges for open science
98.	Giofre (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines
99.	Giofre (2022)	Open science policies and guidelines	Rewards and incentives	badges for open science
100.	Gkiouras (2022)	Open methodology	Open protocol	publication of protocol
101.	Goldacre (2019)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal requires reporting checklist
102.	Goldacre (2019)	Other	Other	discrepancy review
103.	Gorman (2020)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
104.	Gregori (2013)	Other	Type1/Type 2 error reduction	effect-size filter
105.	Gruber (2016)	Other	Type1/Type 2 error reduction	p-value calibration
106.	Gruber (2016)	Other	Type1/Type 2 error reduction	p-value calibration
107.	Hamilton (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines
108.	Hamilton (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data/code sharing policy
109.	Hamilton (2023)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines

110.	Hardt (2013)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE proposed trial registration
111.	Hardt (2013)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal policy on trial registration
112.	Hardwicke (2018)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal open data policy
113.	Hardwicke (2021)	Open science policies and guidelines	Rewards and incentives	badges for open data
114.	Haven (2023)	Other	Other	role models
115.	Hopewell (2012)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal endorses reporting checklist
116.	Hopewell (2016)	Open science tools	Other	writing tool
117.	Howell (2015)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
118.	Jaljuli (2023)	Other	Other	statistical approach
119.	Kidwell (2016)	Open science policies and guidelines	Rewards and incentives	badges for open data and materials
120.	Krawczyk (2012)	Open data and materials	Affirmative sharing declarations	data sharing statements
121.	Ladd (2023)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines
122.	Ladd (2023)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal submission guidelines
123.	Laurinavichyute (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal open data policy
124.	Laurinavichyute (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal open data policy
125.	Liu (2022)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
126.	Madduri (2019)	Open science tools	Open workflow tools, workflow management systems	workflow tools
127.	Magee (2014)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
128.	Marin (2015)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	FDA trial registration policy
129.	McGee (2011)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal policy on trial registration
130.	Menon (2021)	Open educational resources	Training	training

131.	Meredith (2019)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
132.	Milia (2012)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
133.	Narang (2023)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE data sharing statement
134.	Naudet (2018)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
135.	Naudet (2018)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
136.	Nguyen (2022)	Open science policies and guidelines	Reporting guidelines, reporting standards	use of reporting checklist
137.	Nguyen (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
138.	Nuijten (2017)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
139.	Nuijten (2017)	Open science policies and guidelines	Rewards and incentives	badges for open science
140.	Nuijten (2017)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
141.	Obels (2020)	Open methodology	Registered reports	registered reports
142.	Palmer (2021)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
143.	Piwowar (2010)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal and funder policies
144.	Piwowar (2011)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	funder data sharing policy
145.	Piwowar (2011)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
146.	Riedel (2022)	Other	Collaborative research	building a (training) community
147.	Rowhani-Farid (2016)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
148.	Rowhani-Farid (2018)	Open science policies and guidelines	Rewards and incentives	badges for data sharing
149.	Rowhani-Farid (2018)	Open science policies and guidelines	Rewards and incentives	badges for code sharing
150.	Rowhani-Farid (2020)	Open science policies and guidelines	Rewards and incentives	badges for open data

151.	Ruamviboonsuk (2023)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE data sharing statement
152.	Salameh (2019)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
153.	Savage (2009)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
154.	Shamliyan (2014)	Open methodology	Pre-registration, study registration, analysis plan	trial registration
155.	Shen (2023)	Open methodology	Project workflow	standardization of processes
156.	Siebert (2020)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE data sharing statement
157.	Sims (2016)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	ICMJE proposed trial registration
158.	Sims (2018)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal policy on trial registration
159.	Sinaci (2023)	Open science tools	Other	standardization of processes
160.	Soderberg (2021)	Open methodology	Registered reports	registered reports
161.	Speich (2023)	Other	Other	editorial checks/peer review
162.	Struthers (2021)	Open science tools	Other	writing tool
163.	Thelwall (2017)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal data sharing policy
164.	Toth (2021)	Open methodology	Pre-registration, study registration, analysis plan	preregistration
165.	Toth (2021)	Open methodology	Pre-registration, study registration, analysis plan	preregistration
166.	Toth (2021)	Open methodology	Pre-registration, study registration, analysis plan	preregistration
167.	Trisovic (2022)	Open science tools	Other	automatic code cleaning
168.	Trisovic (2022)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal policy strictness
169.	van der Braak (2022)	Open methodology	Pre-registration, study registration, analysis plan	preregistration
170.	Vassar (2020)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
171.	Veroniki (2021)	Open science policies and guidelines	Reporting guidelines, reporting standards	publication of reporting checklist
172.	Walsh (2020)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal endorses reporting checklist
173.	Witwer (2013)	Open science policies and guidelines	Policy guidelines (e.g., of funders/publishers)	journal requires reporting checklist

B. Score sheet and explanations

This Table presents the instructions that were given to the assessors.

	ITEM	DESCRIPTION	EXPLANATION
BASIC DETAILS	Intervention	Define the intervention of interest, including who implements the intervention and who is affected.	Note: we may have no studies that look at exactly this situation; sometimes the institute or individual who will do the implementation may differ between studies; or they may be different from the implementor who is consulting the toolbox. Sometimes, there may be some variation in the intervention itself (e.g. we combined "instructions for authors" with "multiple journal editors recommended"). In this case, the implementers in the studies we look at should still be considered, but the mismatch can be considered under 'applicability' or 'inconsistency' (see below).
	Outcome	In principle, we will consider "reproducibility" as enabling others to re-do your study and thus require, request, encourage to be as transparent as possible.	Some studies may focus on other aspects of reproducibility and replicability, such as: 1. whether the study can actually be reproduced when re-done; 2. whether data and code are shared; 3. transparent reporting
	Number of studies	How many relevant studies were included in the (scoping) review that address this intervention?	Please consider and explain whether these studies addressed exactly the same intervention, or whether the studies differed on crucial points: how many publications study an intervention that falls into the same type? If a (recent) systematic review of meta-analysis was available, then only one report (the review) can be sufficient.
	Meta-analysis or systematic review available?	yes or no	If a meta-analysis was among those studies, then it will not be necessary to separately assess the individual studies as well. Note that a systematic review may be without a meta-analysis; and note that the meta-analysis needs to be sufficiently recent and complete.
	Conclusions	What were the authors' conclusions about the effectiveness of the intervention?	If all relevant studies point in the same direction, then we can use one answer for all studies. If the studies differ, then please explain if a "majority vote" was chosen and whether another choice was made.
	Quantitative results	Was there a relevant quantitative result presented, indicating the effect of the intervention?	For most interventions, the quantitative results will be too variable between studies or there will be too many results presented for sub-categories of the outcome. Only assess this if a meta-analysis for a relevant outcome is available.

BASIC DETAILS	Comparative experimental design	Did all or some of the studies use a comparative experimental design?	Has the intervention been assessed in at least one study with an experimental design (the researchers decide who receives the intervention and who not) in which the effect of the intervention was compared against the effect of another or no intervention?
	Comparative studies	The majority of the studies compared the effects of an intervention against the situation where no or another intervention was used	Has the intervention been assessed in at least one study with any comparative (either experimental or observational) design in which the effect of the intervention was compared against the effect of another or no intervention?
	Type of comparative study	What kind of comparative studies were included?	• Independent before-after study = a study in which two different groups were compared; one group before implementation and one group after implementation (e.g. a group of studies before the publication of the CONSORT checklist compared with a group of studies published after the publication of the CONSORT checklist). • Parallel arms at the same time: this can be experimental (researchers assign who receives intervention) or observational (two 'naturally occurring' groups). For example, studies published in open access journals compared with studies published in non-open access journals. • Before-after-same-subjects: for example training. For example, percentage of preprints of a group of researchers before training compared with the percentage of preprints from that same group of researchers after training. • N/a: the intervention was applied and then the effects were checked, without a formal comparison with a situation without the intervention.
	Prospective study design	Were most of the studies for this intervention planned before the moment of the baseline measurement?	For example, a study that looked back at an intervention that was implemented two years ago, does not use a prospective design.
	Precision – sample sizes	How many studies address this intervention?	
	Precision – sample sizes	What is the median (range) sample size of the included studies?	Report the median and the range (so NOT the mean and SD).

	Precision – quantitative result	If a quantitative result was presented, how precise was this?	Please describe the estimate plus confidence interval (or other measure of uncertainty). If studies provide quantitative results, but it is not clear how valuable these are for our decision; and neither to what extent they relate to the relevant outcomes. (e.g. separate number for hypothesis-reporting and separate for sample size justification and separate for number of items reported....) then we ignore this one.
	Consistency	Do the results of all studies point into the same direction?	They all five have positive conclusions towards the intervention promoting reproducibility.
	Variation (heterogeneity)	To what extent is the body of evidence diverse and heterogeneous? Are there concerns that the body of evidence is not applicable to our situation?	Different interventions, different fields, different outcome measures.

Under review by the European Commission

C. Assessment sheet

For reasons of documentation and transparency, the tables below present the form, where the assessment opinions of all involved researchers were compiled by Mariska Leeflang.

Assessment Table 1 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	FULL DESCRIPTION	OUTCOME	N studies
1	Infrastructure	Publishers	Journal policy on completeness of reporting	Journals adopt/endorse a policy requiring or encouraging completeness of reporting (e.g. endorsing CONSORT, COREQ, STROBE, PRISMA, ARRIVE, or reminding peer reviewers of guideline items). Implemented by journals/editors/publishers; affects submitting authors and, indirectly, readers.	Improved code and data sharing, transparent reporting.	18
2	Infrastructure	Publishers	Journal policy on data sharing	Journals ask in several ways for authors to share their data: ask for a statement (2 studies); award badges (1 study); state that they adhere to ICMJE policy (1 study); mandate data sharing (11 studies); encourages data sharing (13 studies); mentions it is author instructions (1 study); requests data from authors (1 study)	Increasing data sharing behaviour (22); Reproducibility (4); Reusability (2); Inclusion of data sharing statement (3); Reporting inconsistencies (1); Excess statistical significance (1)	29
3	Infrastructure	Publishers	Journal policy on registration	Journals ask from researchers to register their RCTs.	Fostering transparency, replicability and redoability	7
4	Infrastructure	Publishers	Journal awards open science badges	Journals award badges to authors when they make their data, code or materials open and publicly accessible.	Increasing data sharing behaviour (5); Computational reproducibility (2); reporting inconsistencies (1)	8
5	Infrastructure	Publishers	Results-free peer-review	Journals implement a policy where authors submit their manuscripts to the journal devoid of empirical results and then journals reviewing the	Removement of bias in editor/reviewer decisions	1

				manuscripts through to an accept-reject decision without ever seeing the results.	(but they do not formally assess an outcome)	
6	Infrastructure	Publishers	Outcome-agnostic acceptance by journal	Editors to encourage authors and guide referees of eight health economics journals with an editorial statement to make their submissions (authors) or decisions (referees) on the potential scientific and publication merit regardless of whether studies' empirical findings do or do not reject null hypotheses.	To decrease the extent of publication bias.	1
7	Infrastructure	RFOs	Funder policy on data management	Research funding organisation require researchers to submit a data sharing plan	Increasing data sharing behaviour	3
8	Infrastructure	RPOs	Research institution provides support services	Research organisations provide institutional infrastructures, policy directives, and automated tools to guide and assist researchers in making their data FAIR throughout research process or lifecycle.	Whether a study can be reproduced	2
9	Infrastructure	RPOs	Research institution provides training	This can be any training around reproducibility; the included studies looked at (1) institute provides workshops about data management; (2) funding organisation provides training on systematic reviewing; (3) institute arranged programing workshop.	Reproducibility rate; quality and transparency of research; knowledge and attitude	3
10	Infrastructure	RPOs	Research institution provides role models	Encourage PhD supervisors to engage in Open Science practices to act as role models to their supervisees and in this way encourage them to also engage in Open Science practices.	Open Access publication	1
11	Infrastructure	Other	Government policy on registration	Government implements legislation on trial registration, aiming to increase availability of study results	Availability of results	2
12	Infrastructure	Other	Data repository	Institute, organisation, government, publisher or funder set up repository; or mandate use of repository	Data sharing and re-execution of studies; output consistency	3

13	Preparation	Formulate the RQ	N/A; not available in review(s)	Explicit and transparent formulation of the research questions, for example by a formalized and open process in which stakeholders gather together to formulate gaps in knowledge.	N/A	0
14	Preparation	Plan the study	Preregistration	Researcher publishes a protocol of their study (before the results are known)	Transparent reporting (2/3); computational reproducibility (1); rigour (2)	6
15	Preparation	Plan the study	Optimization of study design	A wide varying range of interventions: (1) standardization of effect sizes; (2) best practices for western blot analyses; (3) Python toolbox that standardizes experiments; (4) association of effect size with reproducibility.	Improve their rigor and reproducibility of western blot experiments (Omondi). Improve reproducibility in behavioral VR experiments (Schuetz). Promote replicability (Li).	3 (It was 5, but one is preprint and was removed and there is one duplicate)
16	Preparation	Plan the study	Statistical analysis plan	A statistical Framework that outlines a systematic approach for causal inference analyses	pre-specification of the data analysis plan	1
17	Preparation	Plan the study	Standardization of experimental protocol	Researchers develop a harmonized research protocol	Reproducibility in preclinical studies	2
18	Preparation	Obtain resources	Researcher trained on study protocol	Researchers receive training in reproducibility aspects	Varying outcomes	3 (but only 1 addresses only training; the other two studies also introduce a structured workflow)
19	Data collection	Collect data	Researcher uses standard method for data acquisition	Research teams develop, implement and evaluate standardized physical protocols, harmonized experimental procedures, and quality control frameworks to improve data reproducibility	Inter-laboratory variability and reproducibility of experimental findings across independent research sites.	11

20	Data collection	Collect data	Multi-site testing	EXCLUDE		1
21	Data collection	Store data	Standardized file system structure	Researchers use a standardized, pre-defined structure to organize data according to FAIR principles	Whether the study can actually be reproduced when re-done	3
22	Data collection	Store data	Data storage solution	Researchers use repository to store/share data. OR in general: existence of a repository to store/share data (if we focus more on the infrastructure itself)	Promote the sharing and interoperability of data, with the added benefit of safe storing/sharing.	3
23	Data collection	Store data	Direct data deposition	Researchers directly deposit incoming data to a public repository	Availability of data	1
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework	EXCLUDE		1
25	Data collection	Document data collection procedures	Software tool that documents data retrieval procedures	A variety of different interventions, such as automated data curation in R; and standardized mass-spectrometry measurements.	Improved reproducibility, transparency and/or reliability	2
26	Data collection	Document data collection procedures	Real-time procedure documentation	A bundle of credibility-enhancing research methods: (1) an expert-consensus-designed preregistered protocol; (2) open data and automated, preregistered analysis code; (3) blinding and other bias-control procedures; (4) independent external research audits during data collection.	Improved reproducibility as a methodology by eliminating possibilities of questionable research practices.	1
27	Data analysis	Prepare data	Standardized data preparation procedure	A wide varying range of interventions: (1) harmonisation of self-reported sleep data; (2) automated preparation of MRI data; (3) automated report on data properties prior to analysis; (4) standardization of proteomics	Replicability, reliability and generalizability; Data uniformity	5

				analyses; (5) normalization of metabolomics data.		
28	Data analysis	Prepare data	Data quality control	Biologists and biomedical researchers evaluate biological sample outliers for data quality control by addressing quality imbalance across experimental groups	The paper argues that quality issues can undermine reproducibility, but this is inferred and not tested.	1
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	Coordinating research centres and software developers mandating shared standardized analytical algorithms to process data	Independent labs reach identical conclusions, code and data sharing via open source pipelines to allow re-execution and measurement standardisation	19 (but only 3 were proper evaluations)
30	Data analysis	Conduct analyses	Statistical tool to estimate power	IS THIS REALLY AN INTERVENTION?		0
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	software tool development and infrastructural engineering	rigor, statistical sensitivity, reproducibility and/or replicability	2
32	Data analysis	Conduct analyses	Bias analysis method	IS THIS REALLY AN INTERVENTION?		0
33	Data analysis	Validate analysis	Multiverse analysis	Implementation of multiverse analyses to assess the robustness of research results to analytic decisions.	Authors, reviewers, and editors to incorporate multiverse analyses to enhance replicability of research findings	1
34	Data analysis	Validate analysis	Authors use effect-size filter	Researchers use statistical improvement method (?) Specifically: Adjusting p-values to account for random and systematic errors (Gruber); Use of post-test filter based on effect size and signal level thresholds (Gregori); Use of partial-pooling analyses (Senior)	Replicability (obtaining similar/comparable results across laboratories)	3

35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	Different interventions that either standardize analyses; or provide a codebook-like way of documenting them	Reproducibility	2
36	Data analysis	Document data analysis procedures	Open workflow platform	Researchers use a web-based or cloud-hosted environment to build, execute and/or share computational pipelines	Whether the study can actually be reproduced when re-done	3 (but they are all different tools)
37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	Researchers use tool to standardise reporting and data analysis frameworks and deliver reliable, FAIR compliant and reproducible research data	Successful delivery of reproducible data and Transparent reporting	4
38	Dissemination	Write the report	Use of reporting checklist	The use of a reporting checklist or writing aid by individual researchers	completeness, quality, or reproducibility of research reporting	15
39	Dissemination	Share data	Researchers share data	Researchers (state they) make their data available (upon request)	Research reproducibility and whether data and code are shared	4
40	Dissemination	Share data	Data management tool	Five distinct software/infrastructure tools that address different aspects of reproducibility: platform to analyze datasets without publicly sharing data; enrichment of FAIR data with provenance-related metadata; automatic FAIRness assessment; community-driven blockchain-based platform.	reproducibility, FAIR compliance, traceability, or reusability of research data/code/experiments in its domain	5
41	Dissemination	Peer-review	Peer-reviewer checks reporting	Journal asks peer reviewers to check reporting or mentions reporting standards to encourage reporting from authors or editors check themselves.	Completeness of reporting	4
42	Dissemination	Post-publication	Post-publication data requests	Researchers request data from authors after publication	Availability of data	2

43	Dissemination	Post-publication	Discrepancy review	Researchers do a discrepancy review and send correction letters to journal editors.	Raw data shared in real time. The trials' misreporting was made completely transparent.	1
44	Dissemination	Post-publication	Replication	Researchers replicate or reproduce someone else's projects	I wonder if we should keep this. This is about interventions at the side of people who are actually reproducing studies.	3

Assessment Table 2 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	SR or MA?*	CONCLUSIONS	QUANTITATIVE RESULTS (of MA was conducted)
1	Infrastructure	Publishers	Journal policy on completeness of reporting	1 systematic review of trials	Positive effect	no
2	Infrastructure	Publishers	Journal policy on data sharing	1 SR plus MA (Hamilton 2023)	14 positive effect; 16 neutral or no effect	105 meta-research studies examining 2 121 580 articles across 31 specialties: 8 had low risk of bias; outcome = historical trends in data sharing (statements and actual).
3	Infrastructure	Publishers	Journal policy on registration	no	Positive effect	No
4	Infrastructure	Publishers	Journal awards open science badges	no	Positive effect	No
5	Infrastructure	Publishers	Results-free peer-review	no	Positive	no
6	Infrastructure	Publishers	Outcome-agnostic	no	Positive effect	No

			acceptance by journal			
7	Infrastructure	RFOs	Funder policy on data management	no	No effect	No
8	Infrastructure	RPOs	Research institution provides support services	no	Positive effect	no
9	Infrastructure	RPOs	Research institution provides training	no	Positive effect (Reidel & Menon). No effect (Deardorff).	No
10	Infrastructure	RPOs	Research institution provides role models	No	Unclear	N/A
11	Infrastructure	Other	Government policy on registration	No	Effective/unclear	No
12	Infrastructure	Other	Data repository	no	positive effect	no
13	Preparation	Formulate the RQ	N/A; not available in review(s)			
14	Preparation	Plan the study	Preregistration	No. However, multi-lab data were pooled and analyzed, but this is more a single large, coordinated experiment.	Positive effect	No

15	Preparation	Plan the study	Optimization of study design		Positive or unclear effect	No
16	Preparation	Plan the study	Statistical analysis plan	No	Positive effect	No
17	Preparation	Plan the study	Standardization of experimental protocol	Yes, one SR; but the SR is mainly descriptive	Positive	No
18	Preparation	Obtain resources	Researcher trained on study protocol	None	Conclusions are positive; but not really tested	N/A
19	Data collection	Collect data	Researcher uses standard method for data acquisition	No	positive effect	N/A
20	Data collection	Collect data	Multi-site testing			
21	Data collection	Store data	Standardized file system structure	No	Positive effect (but not quantitatively supported)	No
22	Data collection	Store data	Data storage solution	no	Positive effect	No
23	Data collection	Store data	Direct data deposition	No	Unclear	No
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework			
25	Data collection	Document data collection procedures	Software tool that documents	N/A	Positive effect	N/A

			data retrieval procedures			
26	Data collection	Document data collection procedures	Real-time procedure documentation	no	positive effect (only 1 study)	no
27	Data analysis	Prepare data	Standardized data preparation procedure	No	Positive	N/A
28	Data analysis	Prepare data	Data quality control	No	Positive effect	No
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	No	positive (for 13 out of 19 studies)	N/A
30	Data analysis	Conduct analyses	Statistical tool to estimate power			
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	N/A	Positive effect	N/A
32	Data analysis	Conduct analyses	Bias analysis method			
33	Data analysis	Validate analysis	Multiverse analysis	no		No
34	Data analysis	Validate analysis	Authors use effect-size filter	No	Effective	No
35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	None	Both positive, but based on anecdotal evidence	N/A

36	Data analysis	Document data analysis procedures	Open workflow platform	No	Effective	No
37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	No	Positive effect	N/A
38	Dissemination	Write the report	Use of reporting checklist	No	Overall effective, but weak effects	N/A
39	Dissemination	Share data	Researchers share data	no	Positive effect	no
40	Dissemination	Share data	Data management tool	No	All positive	N/A
41	Dissemination	Peer-review	Peer-reviewer checks reporting	No	Positive effect	N/A
42	Dissemination	Post-publication	Post-publication data requests	no	Unclear, but some words (e.g. mentioning replication crisis) led to more data shared	The studies present sharing rates in terms of percentages
43	Dissemination	Post-publication	Discrepancy review	No	Positive effect	N/A
44	Dissemination	Post-publication	Replication	None	?	N/A

Assessment Table 3 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	COMPARATIVE EXPERIMENTAL DESIGN	COMPARATIVE STUDIES	TYPE OF COMPARATIVE STUDY	PROSPECTIVE STUDY DESIGN
-----	----------	-------------	------	---------------------------------	---------------------	---------------------------	--------------------------

1	Infrastructure	Publishers	Journal policy on completeness of reporting	Yes (n=3)	Yes	mix – mostly ☒ Independent before-after (endorsement audits, ITS such as Hopewell 2012) PLUS ☒ parallel arms (endorsing vs non-endorsing journals) and the RCTs (Speich 2023).	Yes
2	Infrastructure	Publishers	Journal policy on data sharing	No	Yes, 14 comparative studies	most cross-sectional; 3 pre-post design; 5 longitudinal	None
3	Infrastructure	Publishers	Journal policy on registration	No	Some	Parallel arms	Some
4	Infrastructure	Publishers	Journal awards open science badges	yes - 1 study	Yes	Independent before-after study & parallel arms	One
5	Infrastructure	Publishers	Results-free peer-review	no	no	N/A	Yes (but they do not make a comparison)
6	Infrastructure	Publishers	Outcome-agnostic acceptance by journal	Yes	Yes	Parallel arms observational AND Independent before-after	No
7	Infrastructure	RFOs	Funder policy on data management	No	Yes	Parallel arms	No
8	Infrastructure	RPOs	Research institution provides support services	no	no	no	no
9	Infrastructure	RPOs	Research institution provides training	Yes	Yes (2 in out of 3 studies).	Independent before after (Reidel). Before after same subject (Deardorff).	No (Yes to only Deardorff).

10	Infrastructure	RPOs	Research institution provides role models	No, observational study	Yes	Independent groups comparison	No
11	Infrastructure	Other	Government policy on registration	No	One	Independent before-after study	No
12	Infrastructure	Other	Data repository	no	no	no	no
13	Preparation	Formulate the RQ	N/A; not available in review(s)				
14	Preparation	Plan the study	Preregistration	No		Parallel arms	No
15	Preparation	Plan the study	Optimization of study design	Yes (only Omondi).	Yes	Parallel arms experimental (Omondi).	Yes (Omondi)
16	Preparation	Plan the study	Statistical analysis plan	No	No	N/A	No
17	Preparation	Plan the study	Standardization of experimental protocol	No	Yes	Observational before-after	No
18	Preparation	Obtain resources	Researcher trained on study protocol	None	None	N/A	Yes, 1/3
19	Data collection	Collect data	Researcher uses standard method for data acquisition	Mixed, but mostly not randomized. Only study 9 consisted of controlled mock sample experiment (that comes close to randomization)	Yes for all 11 studies	Parallel arms AND before-after study with the same subjects	Yes

20	Data collection	Collect data	Multi-site testing				
21	Data collection	Store data	Standardized file system structure				No
22	Data collection	Store data	Data storage solution	No	No	N/A	No
23	Data collection	Store data	Direct data deposition	No	No	NA	No
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework				
25	Data collection	Document data collection procedures	Software tool that documents data retrieval procedures	NO	Not comparative studies in the sense that these interventions have been compared to other interventions to check which intervention leads more to reproducibility (these are more like diagnostic studies)	Identical underlying data or samples were analysed twice (once with the standard/manual approach, once with the new tool),	No
26	Data collection	Document data collection procedures	Real-time procedure documentation	No	Unclear	replication comparison against the original 2011 BEM study	Yes

27	Data analysis	Prepare data	Standardized data preparation procedure	Yes (only Ashkarran et al. and Tang et al.)	Yes (2 out of 5 studies)	Independent before-after, same subjects (Ashkarran et al. and Tang et al.).	None
28	Data analysis	Prepare data	Data quality control	Yes	Yes	Before-after with the same subjects before and after the intervention (they look at the same datasets, and see what happens when you do address QI versus when you do not)	No
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	No	Yes	1 before-after study; 1 parallel arms study	some
30	Data analysis	Conduct analyses	Statistical tool to estimate power				
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	No	Yes	Closest to before-after / same-subjects for both:	Yes for both, in different ways
32	Data analysis	Conduct analyses	Bias analysis method				
33	Data analysis	Validate analysis	Multiverse analysis	Yes	Yes	Experimental parallel arms	No
34	Data analysis	Validate analysis	Authors use effect-size filter	Yes	Yes	Before-after study with the same subjects before and after the intervention	No
35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	None	None	N/A	None

36	Data analysis	Document data analysis procedures	Open workflow platform	No	No	NA	No
37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	No	No	N/A	No
38	Dissemination	Write the report	Use of reporting checklist	Yes: more than 1 study: 2 did, and one had a quasi-experimental comparative design	Yes	Mixed: independent before-after in most.	Mixed. Several are prospective, most are retrospective audits
39	Dissemination	Share data	Researchers share data	no	no	NA	Most retrospective studies
40	Dissemination	Share data	Data management tool	No	limited: 2 comparative studies; one partially; 2 feasibility studies, uncompar	When applicable: before-after in same subject	Mixed: 1 and 3: prospectively, other retrospectively
41	Dissemination	Peer-review	Peer-reviewer checks reporting	Yes (3 RCTs)	Yes	3 RCTs; 1 before-after	Yes
42	Dissemination	Post-publication	Post-publication data requests	Yes	Yes	Parallel arms at the same time	Yes
43	Dissemination	Post-publication	Discrepancy review	No	No / unclear	N/A	Yes
44	Dissemination	Post-publication	Replication	None	None	N/A	None

Assessment Table 4 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	PRECISE (STUDIES)	PRECISE (OBSERVATIONS)	PRECISE (CONFIDENCE INTERVAL)	CONSISTENCY
1	Infrastructure	Publishers	Journal policy on completeness of reporting	More than 10	4-115	N/A	little consistency
2	Infrastructure	Publishers	Journal policy on data sharing	30 studies, including one SR		N/A	Relatively consistent, as all studies are either positive or neutral
3	Infrastructure	Publishers	Journal policy on registration	5 to 10	Studies include varying units of analysis	N/A	No
4	Infrastructure	Publishers	Journal awards open science badges	5 to 10	N/A (different units of analysis)	N/A	Yes
5	Infrastructure	Publishers	Results-free peer-review	<5	19 submitted studies	N/A	N/A
6	Infrastructure	Publishers	Outcome-agnostic acceptance by journal	<5	Total of 230 publications with editorial statement.	The intervention decreased the proportion of tests rejecting the null hypothesis by 18%.	Yes
7	Infrastructure	RFOs	Funder policy on data management	<5	397 (213-11,603)	Unclear	Yes
8	Infrastructure	RPOs	Research institution provides support services	<5	varying unit of analysis	N/A	yes

9	Infrastructure	RPOs	Research institution provides training	<5	Different units of analysis. 108 publications; 54 researchers; 14 interviews	No	No
10	Infrastructure	RPOs	Research institution provides role models	<5	211	Not very precise	Yes
11	Infrastructure	Other	Government policy on registration	<5	234-12,938	Precise/unclear (increase in proportion of published studies from 64% to 85% ($p = 0.032$; proportion of reported results from 14% to 70% $p < .001$)	Yes
12	Infrastructure	Other	Data repository	<5	1 study presented 86 implemnetations; other two do not report numbers	unclear	yes
13	Preparation	Formulate the RQ	N/A; not available in review(s)				
14	Preparation	Plan the study	Preregistration	5 to 10	222 (1-2127)	N/A	Yes
15	Preparation	Plan the study	Optimization of study design	<5	Harnessing a database of 6000 + western blots from N=281 subjects (Omondi). 5 people that tested the VR experiment. 307 original-	Unclear (Omondi, Schuetz).	No

					replication study pairs (Li).		
16	Preparation	Plan the study	Statistical analysis plan	<5	N/A	N/A	N/A
17	Preparation	Plan the study	Standardization of experimental protocol	<5	3 labs, 569 papers	N/A	They point in the same direction (positive), but they look at different outcomes
18	Preparation	Obtain resources	Researcher trained on study protocol	3	1 study presented 14 interviews; other two do not report numbers	N/A	difficult to say - two studies do not really investigate training effect
19	Data collection	Collect data	Researcher uses standard method for data acquisition	>10	Unclear; very diverse	Only few CI reported	Mostly consistent in direction (standardization improves reproducibility)
20	Data collection	Collect data	Multi-site testing				
21	Data collection	Store data	Standardized file system structure	<5	Not very meaningful (ENCORE: 20 projects, FHIR: 2 institutions, LASSO: 158 coding problems)	Imprecise/NA	Yes (all conclude positively, but the exact interventions are different)
22	Data collection	Store data	Data storage solution	<5	1 (1)	Unclear	Yes

23	Data collection	Store data	Direct data deposition	<5	1 case study	NA	NA
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework				
25	Data collection	Document data collection procedures	Software tool that documents data retrieval procedures	n=2	Case study 1 compared 88,782 vs 86,705 records (98% overlap in magnitude); case study 3 retrieved 177,634 records; study 3 4600-4800 (global benchmark) to ~6500 (melanoma cohort) protein groups identified across 30-95 patient samples	N/A	Yes
26	Data collection	Document data collection procedures	Real-time procedure documentation	<5	2115 participants in primary analysis, 10 lab sites	No confidence interval	N/A
27	Data analysis	Prepare data	Standardized data preparation procedure	5	Varying units of analysis.	N/A	Yes
28	Data analysis	Prepare data	Data quality control	<5	40 datasets	Imprecise	N/A only one study
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	>10	20- 55273	unclear	All positive

30	Data analysis	Conduct analyses	Statistical tool to estimate power				
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	<5	Very precise and large for brainlife.io: >3,200 participants across HCP (1,066 + 44 test-retest), Cam-CAN (652), PING (110), ABCD (1,877), HBN (42), plus smaller clinical cohorts (Indiana concussion N=29; Oxford eye disease dataset). blotRig: a database of 6,000+ western blots from N=281 subjects (rats and mice) used to build/validate the tool, with the worked demonstration on N=29 rats (3 technical replicates each).	Precise, with real inferential statistics in both.	Yes
32	Data analysis	Conduct analyses	Bias analysis method				
33	Data analysis	Validate analysis	Multiverse analysis	<5	Five cycles of NHANES with the	Mixed, in some cases precise in some cases not precise.	Yes

					participation of 25,962 adults in total.		
34	Data analysis	Validate analysis	Authors use effect-size filter	<5	NA	Unclear	Yes
35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	2	No real comparison	N/A	N/A
36	Data analysis	Document data analysis procedures	Open workflow platform	<5	Madduri: 11 human users, Bomtempi: 2 use cases, Martinc: 1 use case	Imprecise: Madduri showed that 10/11 could reproduce the analyses; besides this, it's mainly just showing usefulness	Yes/NA (within their own tool, they all point towards a positive effect)
37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	<5	PROTEOMAS: 25 datasets; other 2 studies multiple case studies/data	No CI	Yes
38	Dissemination	Write the report	Use of reporting checklist	>10	Different units of analysis; difficult to indicate whether effects are precise or not	N/A	Yes
39	Dissemination	Share data	Researchers share data	<5	83 (1-2009)	N/A	yes
40	Dissemination	Share data	Data management tool	5-10	Different units of analysis	N/A	Yes
41	Dissemination	Peer-review	Peer-reviewer checks reporting	<5	24-955	N/A	Yes

42	Dissemination	Post-publication	Post-publication data requests	<5	1 study with 200 emails, 1 study with 1028 emails	1 study Highly Precise (95% CI: 34.5%, 40.5%), and the other moderately precise (CI: 37.1%, 50.9%)	Yes
43	Dissemination	Post-publication	Discrepancy review	<5	58-67	precise	N/A
44	Dissemination	Post-publication	Replication	3	2 studies tried to predict replication success	N/A	Ddifferent designs

Assessment Table 5 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	Heterogeneity	Notes	Quality judgment
1	Infrastructure	Publishers	Journal policy on completeness of reporting	Yes, different fields, journals and people	Relatively important evidence for reminders and endorsements... but weak overall. Results diverge based on how actively or passively the journals enforced their policies.	Relatively high
2	Infrastructure	Publishers	Journal policy on data sharing	Variation in exact intervention, in study type, in outcomes		
3	Infrastructure	Publishers	Journal policy on registration	Yes	Four of five retrieved studies had a comparative design All studies conclude that preregistration improves reproducibility No experimental design Only 5 heterogeneous studies	
4	Infrastructure	Publishers	Journal awards open science badges	Yes	The majority of studies measuring sharing behaviour report an increase in data sharing behaviour. Studies that measure re-executability show failure rates, and the RCT shows no effect.	

5	Infrastructure	Publishers	Results-free peer-review	No	Only 1 study, being a pilot study, tht does not make a comparison. Effectiveness is hard to judge, but they do evaluate based on the experiences of reviewers and editors, which gives an indication of feasibility	
6	Infrastructure	Publishers	Outcome-agnostic acceptance by journal	No	A study that followed two designs, i.e. parallel arms observational AND Independent before-after in a rather small (my subjective opinion) group of publications at the specific research field of health economics.	The results seem reliable.
7	Infrastructure	RFOs	Funder policy on data management	No	The studies show similar conclusions: the NIH policy does not seem very effective. No experimental designs Only 3 studies; 1 without comparison Only NIH policy assessed	Studies investigated NIH policy and found no effect. Low certainty.
8	Infrastructure	RPOs	Research institution provides support services	High heterogeneity - two studies with two very different interventions	Evidence is based on only two studies on two very different types of interventions (research data management in institution and automated checks for FAIRness). Interventions are ot really tested; wet-lab and clinical trials only.	Two very different non-comparative studies show a positive effect. Low certainty.
9	Infrastructure	RPOs	Research institution provides training	Yes		Reidel uses a very small cohort of studies. Menon declares that the intervention seems promising but findings need to be interpreted within the narrow scope of the case study. Deardorff uses a small cohort of interviewed researchers.

10	Infrastructure	RPOs	Research institution provides role models	No	Strengths worth highlighting: preregistered design with an a priori power calculation; two independent open-data measures ; a supplementary retraction analysis (found only 2 retractions among 81,091 linked publications, both apparently honest errors, none within their own included dataset) as an indirect check for irresponsible-behaviour bias; full data and code openly available on GitHub/Software Heritage - the study itself practices what it studies. Compared internationally: their open-data rate (8.8%) closely matches Serghiou et al. 2021's PubMed-Central-wide estimate (8.9%), while their open-access rate (82.8%) is notably higher than international estimates (Piwowar 2018: ~30% in biomedicine), which the authors attribute to strong Dutch national open-access policy and publisher deals rather than anything about supervisor role-modeling specifically	The results seem reliable.
11	Infrastructure	Other	Government policy on registration	Yes	While one study concludes positively and one negatively, only the positively concluding study made a comparison between before/after the intervention. The study that did not make a comparison had a very complete sample, the study that did make a comparison only looked at a specific research context.	Only two studies, of which one was comparative. Tvery low quality
12	Infrastructure	Other	Data repository	yes	Applicability depends mostly on researchers whether they can maintain long term code and data repostories	Low quality; no real comparative studies.
13	Preparation	Formulate the RQ	N/A; not available in review(s)			
14	Preparation	Plan the study	Preregistration	Yes	The results are not overly positive, and the studies assess different outcomes, but they do all conclude positively towards preregistration/	

					registered reports/ publishing a research protocol. 3 studies are about preregistration; 2 about registered reports; 1 about registering review protocols. 1 is a case study, the others compare studies that are/are not registered	
15	Preparation	Plan the study	Optimization of study design	Variation in exact intervention, in study type, in outcomes	The blotRig methodology seems to be a sound methodology to standardise western blot analysis and increase its inferential reproducibility (Omondi). The vexptoolbox is supported by little supportive data (Schuetz). One of the studies on the list was a duplicate and one preprint was removed. (so from 5 to 3 articles)	Do not really provide an answer
16	Preparation	Plan the study	Statistical analysis plan	N/A	It is a framework for different studies. Very helpful to help other improve their rigour and reproducibility, and could be an intervention, but that is not tested	Very low, as it is only one study; no interventions were empirically tested
17	Preparation	Plan the study	Standardization of experimental protocol	The studies are done within specific contexts, but could apply to others		
18	Preparation	Obtain resources	Researcher trained on study protocol	Different types of training, other actions, and study designs	I wonder if all these studies really fit under this intervention; also, the intervention says training on study protocol, but training is on all sorts of things, but not on writing the protocol...	Low quality; no real comparative studies.
19	Data collection	Collect data	Researcher uses standard method for data acquisition	Very high, eleven different domains, and very specific	Unanimous direction of effect and tested protocols focused on feasible real world fixes	Solid, mostly multicenter and multilab studies
20	Data collection	Collect data	Multi-site testing			

21	Data collection	Store data	Standardized file system structure	Yes		
22	Data collection	Store data	Data storage solution	Yes	All studies conclude positively on the use of repositories No experimental designs Only 3 studies; all different repositories	
23	Data collection	Store data	Direct data deposition	No		
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework			
25	Data collection	Document data collection procedures	Software tool that documents data retrieval procedures	Yes: very different interventions	Well, are these interventions suitable for our toolbox? I would say that is an important question.	Moderate. Both are credible, methodologically transparent single-team papers with real quantitative benchmarking – but both involve self-benchmarking by the tool's own developers.
26	Data collection	Document data collection procedures	Real-time procedure documentation	N/A	This paper is best read as a feasibility/methods-demonstration case study rather than as ordinary intervention evidence	Low certainty - only one study
27	Data analysis	Prepare data	Standardized data preparation procedure	Yes	Two out of five are comparative studies, but with no confidence intervals or reported uncertainties.	Low certainty of evidence; although it seems logical that standardized methods improve reproducibility, these studies do

						not really prove that
28	Data analysis	Prepare data	Data quality control	No, only 1 study in a very specific context		
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	Variation in exact intervention, in study type, in outcomes	Most studies did not do a formal evaluation at all; they just present their standardization method as solution to variation in results. 13 studies points towards common methods increasing reproducibility. 6 show common methods produce divergent results under real world conditions and hence limit or reduce repro.	Low certainty of evidence; although it seems logical that common methods improve reproducibility, these studies do not really prove that
30	Data analysis	Conduct analyses	Statistical tool to estimate power			
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	Yes	Both explicitly target 'hidden' methodological practices that are common but rarely scrutinised. blotRig's worked example is a small but rigorously designed, blinded, prospective experiment with a direct, interpretable statistical demonstration. brainlife.io's evidence base is much larger in scale (>3,200 participants, multiple independent cohorts, several external replications) and is the stronger of the two by sheer weight of validation, though a single N=29 comparison (blotRig) is not intended to generalise the way brainlife.io's multi-cohort replications do - the two are not evidentially equivalent despite both being 'high quality' for their respective scope.	High for both, on different grounds.
32	Data analysis	Conduct analyses	Bias analysis method			

33	Data analysis	Validate analysis	Multiverse analysis	No		
34	Data analysis	Validate analysis	Authors use effect-size filter	Yes		
35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	yes	Comparative setup was used to evaluate the standardised schema against informal baseline documentation rather than only describe the tool in isolation. The studies evaluated real world pipelines rather than theoretical conceptual frameworks.	Low quality; no real comparative studies.
36	Data analysis	Document data analysis procedures	Open workflow platform	Yes (but see note)	Different domains, different platforms. Which makes it heterogeneous in some way, but also not necessarily comparable. As for evidence, no study used a control group, in all studies the people testing the platform were the platform's own developers or people from the developers' own research group	
37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	Very high: different disciplines and methods.	How useful are these tools? And they all use external standards, not creating own standards, right? Methodologies do not show variation, but the studies do show operational heterogeneity. They differ in domains, and the stage in research lifecycle when the intervention is implemented.	Four heterogeneous studies; low certainty
38	Dissemination	Write the report	Use of reporting checklist	Although they all look at reporting guidelines, they all investigate different reporting guidelines	Very difficult to assess the disciplines. But it looks like active endorsements are needed (and not put in practice very often). How helpful can these studies be? Or just the intervention group in general, that shows weak evidence?	Variable
39	Dissemination	Share data	Researchers share data	yes		

40	Dissemination	Share data	Data management tool	Yes	SEE NOTES IN ORIGINAL FILES	Interventions are very heterogeneous; low certainty of evidence
41	Dissemination	Peer-review	Peer-reviewer checks reporting	No	The interventions' success depends on who executes it, as the passive interventions targeting standard peer reviewers showed weak effects whereas active interventions executed by specialised journal experts or policies had strong positive effects. Out of 4 studies, 2 had positive effects and 2 had minimal or no positive effect	High quality, although mixed conclusions about the effectiveness
42	Dissemination	Post-publication	Post-publication data requests	yes		
43	Dissemination	Post-publication	Discrepancy review	N/A	The authors built in their own quality control: each trial was coded by one researcher, checked by a second, then replicated in full by a senior academic in a team meeting; all raw coding sheets were shared publicly, inviting scrutiny from trialists and editors. But the question is how good AI can do this in the future to make it much more scalable and efficient, and less resource and time intensive. The study evaluates real world application of reporting standards providing real world practical evidence May not be generalizable to smaller journals since it is focused on a single elite journal	Low certainty: one observational study
44	Dissemination	Post-publication	Replication	yes		Low quality; no real comparative studies.

Assessment Table 6 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	FINANCIAL COSTS	EFFORT AND TIME	REQUIRED KNOWLEDGE	SET-UP AND INFRASTRUCTURE
1	Infrastructure	Publishers	Journal policy on completeness of reporting	not very high. Besides, burden lies with publishers, and they should put more emphasis	Not a lot for publishers, more for researchers, but then it is retrospective after study finalization	low, can be done by admin staff	Medium
2	Infrastructure	Publishers	Journal policy on data sharing	Depending on level of mandating and extent of policy, costs may vary from virtual none (just say something about data sharing in the author instructions) to moderate costs for journal (how to check whether authors really share data).	Authors: costs time and effort to actually share data; Journals / editors / reviewers: costs time and effort to actually check the data sharing; especially if one wants to check if data have been shared in such a way that third parties can use it.	Little	Depending on level of mandating and extent of policy, costs may vary from virtual none (just say something about data sharing in the author instructions) to high (set up repository)
3	Infrastructure	Publishers	Journal policy on registration	Low costs	Low to medium	None	Low
4	Infrastructure	Publishers	Journal awards open science badges	Low	Low	Some (minimal)	Low
5	Infrastructure	Publishers	Results-free peer-review	Low	Low or medium	No	Low

6	Infrastructure	Publishers	Outcome-agnostic acceptance by journal	Low cost	Low effort	None	Low
7	Infrastructure	RFOs	Funder policy on data management	Low or medium costs	low or medium effort, depending on who you ask	None / Yes	Low
8	Infrastructure	RPOs	Research institution provides support services	Medium/High	Low to high	Yes	Medium
9	Infrastructure	RPOs	Research institution provides training	Medium costs (depending on type of training)	Medium effort and time (depending on type of training)	Not from participants; but teachers require knowledge	Medium to high
10	Infrastructure	RPOs	Research institution provides role models	Medium costs	Medium to high	Knowledge required from mentor / role model	Moderate to high
11	Infrastructure	Other	Government policy on registration	Low	Low	No	Low
12	Infrastructure	Other	Data repository	medium or high	medium, but may vary depending of size of repository	yes	high
13	Preparation	Formulate the RQ	N/A; not available in review(s)				
14	Preparation	Plan the study	Preregistration	Low	Low or medium	No	Low
15	Preparation	Plan the study	Optimization of study design	Unclear - too heterogeneous	Unclear - too heterogeneous	Unclear - too heterogeneous	Unclear - too heterogeneous

16	Preparation	Plan the study	Statistical analysis plan	Low	Medium	Yes	Low
17	Preparation	Plan the study	Standardization of experimental protocol	Low	Medium	Yes	Low
18	Preparation	Obtain resources	Researcher trained on study protocol	Requires time and maybe a teacher or training platform	Requires time and maybe a teacher or training platform	No knowledge required from training attendants; knowledge needed from teachers.	May require teaching platform
19	Data collection	Collect data	Researcher uses standard method for data acquisition	Highly variable. Can be low, can be high. Depending on many causes	Moderate to high, but eventually there will be lower efforts if routines are adopted	Diverse: sometimes technician level, sometimes organisation and coordinating team	Ranges from minimal to substantial
20	Data collection	Collect data	Multi-site testing				
21	Data collection	Store data	Standardized file system structure	Low	Medium	Yes	Medium
22	Data collection	Store data	Data storage solution	Low or medium	Medium	Some training required	?
23	Data collection	Store data	Direct data deposition	Low	Medium-high	Yes	Medium
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework				
25	Data collection	Document data collection procedures	Software tool that documents data retrieval procedures	Very different between the two.	Different profiles. ECOTOXr: low-to-medium – mostly automated once installed; requires	Different profiles. ECOTOXr: basic-to-intermediate R programming; database structure knowledge only	Different profiles. ECOTOXr: minimal – any computer running R, ~2 GB free disk space. Hybrid-PRM/DIA:

					basic R competence, database build is a one-off (repeatable) step. Hybrid-PRM/DIA: high — needs assay development (peptide panel design, calibration curves), specialised instrument operation	needed for advanced custom queries. Hybrid-PRM/DIA: advanced mass-spectrometry/proteomics expertise	substantial — dedicated high-resolution MS instrument with an API supporting hybrid acquisition
26	Data collection	Document data collection procedures	Real-time procedure documentation	Medium to high	High	Yes	High
27	Data analysis	Prepare data	Standardized data preparation procedure	Medium to high; depending on the exact tools	Medium to high; depending on the exact tools	May need some training	May need infrastructure; depending on how the standardization is done
28	Data analysis	Prepare data	Data quality control	Low	Low	Yes	Low
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	Depends on level of standardization and mandates. Financial costs may be low.	Depends on level of standardization and mandates. Time and effort may be large, as it asks from all collaborators with and between teams to adhere to the standardization structure, which may be simple to very elaborate.	yes	medium

30	Data analysis	Conduct analyses	Statistical tool to estimate power				
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	Low for blotRig, higher for brainlive, relies on cloud	Low for blot, higher for brainlife	high	Seems to be high and technical
32	Data analysis	Conduct analyses	Bias analysis method				
33	Data analysis	Validate analysis	Multiverse analysis	Medium cost	High	Training in multiple methods of statistical analysis.	Low
34	Data analysis	Validate analysis	Authors use effect-size filter	Low	Low-medium	Yes	Low
35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	Financial costs are not necessarily high: one could invent this on their own computer; but if you want to set it up institution wide, then the costs will be higher for maintainance and probably also problem solving.	Both building and maintaining such a structure takes time; but also the use of such a system may take more time.	Depending on the complexity of such systems	Low to medium; Depending on how big the structure should be
36	Data analysis	Document data analysis procedures	Open workflow platform	Medium	Medium/High	Yes	Medium/High

37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	Low	Low to medium	Yes	Low to medium
38	Dissemination	Write the report	Use of reporting checklist	Low costs	Low	Need familiarity with the checklists to use it properly	Low
39	Dissemination	Share data	Researchers share data	Low	medium/high	yes	Medium (you need a repository)
40	Dissemination	Share data	Data management tool	Variable	Medium to high, depending on exact intervention	Variable and depending on exact intervention	Variable and depending on exact intervention
41	Dissemination	Peer-review	Peer-reviewer checks reporting	Varies, depending on the level of implementation	Takes about 4 hours per manuscript to do this really well	Depending on actual methods	Depending on actual methods
42	Dissemination	Post-publication	Post-publication data requests	Low	Low	No	Low
43	Dissemination	Post-publication	Discrepancy review	medium	High	Yes	Unclear
44	Dissemination	Post-publication	Replication				

Assessment Table 7 from 7

No.	CATEGORY	SUBCATEGORY	TYPE	DISADVANTAGES	ACCEPTABILITY JUDGEMENT	FINAL JUDGEMENT
1	Infrastructure	Publishers	Journal policy on completeness of reporting	Monitoring requires more in depth knowledge. How good is the reporting? How well do they comply with prereg for example.	Established standards promote acceptability. But mandatory compliance checks leading to admin	Good evidence, that is generally weak. Effectiveness can be improved with

				Submission friction for researchers	friction can lower acceptability.	additional measures (such as endorsement or monitoring)
2	Infrastructure	Publishers	Journal policy on data sharing	Authors who may be reluctant to share data, go to other journals		
3	Infrastructure	Publishers	Journal policy on registration	Stealing ideas? But also securing ideas. Lack of registration culture. We miss a mechanism to render registration a good practice than helps complete reporting.		All, but one, of the studies are quite old and, presumably, do not reflect the current situation on registration. It is about registration of trials NOT of research in general, so we do not reach out to most of the end users.
4	Infrastructure	Publishers	Journal awards open science badges	The badge in itself may be seen as a quality indicator, whereas the evidence for studies that have a badge actually being reproducible is limited.	Authors are likely to adopt OS practices as badges provide visible recognition without mandate imposition.	This is an intervention that journals can implement, and that seems to have a positive effect on data sharing. At the same time, the effect of badges on reproducibility seems limited.
5	Infrastructure	Publishers	Results-free peer-review	It may be more difficult to find reviewers (listed by the authors); it may not be as applicable to qualitative research (listed by the authors, I personally do not agree)	Costs and efforts from the peer reviewers' side are low; but for the journal it may be a change in decision making process about whether or not a study should be published.	This is an intervention that journals can implement to reduce the potential effect of publication bias. While the evidence base is limited to one pilot study, the reviewers and editors had positive experiences with this intervention. The effectiveness of the intervention has not been evaluated.
6	Infrastructure	Publishers	Outcome-agnostic	No disadvantage.	Some difficulty for acceptance due to the extent of publication bias	This is an intervention that publishers/editors can implement to reduce

			acceptance by journal		that has penetrated existing research culture.	publication bias. It seems to be easy to implement but it might be less easy to be accepted due to the strong existing habit(?) of publication bias.
7	Infrastructure	RFOs	Funder policy on data management	Concerns of re-identification, of losing originality due to competitiveness of the specific field of research or of relatively strong links to industry.		This is an intervention that funders can implement, but usually without any monitoring mechanisms and then/thus the naturalistic effect is unclear. (without monitoring it is easier but also less effective)
8	Infrastructure	RPOs	Research institution provides support services		Requires infrastructure, staff and commitment. Potentially high costs.	
9	Infrastructure	RPOs	Research institution provides training	There is s need to incentivise researchers to share data (Redel). The data overrepresent postdoctoral researchers in biomedical research (Deardorff). The results are not sufficient for drawing safe conclusions.	Without incentives data sharing will fail to become common practice.	Causal relation cannot be justified. Without incentives data sharing will fail to become common practice.
10	Infrastructure	RPOs	Research institution provides role models	A culture not to share data is prevalent. Open Access is usually expensive and this discourages researchers to publish in OA journals.	Role model; acceptable	
11	Infrastructure	Other	Government policy on registration	Requires government to take action; no longer in the hands of the scientific community.	Although the cost and effort seem to be low, this might require some infrastructure to make it	Based on the studies that exist, it seems like the FDAAA legislation positively impacted the transparency on research

					possible. Also, someone may need to make sure that policies are followed.	results. Still, many results are not shared. Monitoring mechanisms could potentially improve this, but this would require a lot of time and resource investment
12	Infrastructure	Other	Data repository	Lack of maintenance of code repositories and high cloud hosting fees can be a common potential disadvantage	Data repositories are already accepted measures to share data	Evidence is only based on three studies; none of them are comparative. Still, data repositories are already implemented in many places and by different policy levels.
13	Preparation	Formulate the RQ	N/A; not available in review(s)			
14	Preparation	Plan the study	Preregistration	People may feel that someone can "steal" their ideas when they preregister; however one could also argue the other way around	Relatively cheap and effortless; But if referees need to be trained and installed, then the costs and efforts may be much higher	While the body of evidence for preregistration is not overly optimistic or broad, all studies conclude positively towards preregistration. This conclusion is based on various outcomes, including transparent reporting, computational reproducibility and rigour. Given this variety of benefits of preregistration, and its applicability in a multitude of research contexts, this intervention is evaluated positively.
15	Preparation	Plan the study	Optimization of study design	Unclear - too heterogeneous		This type of interventions is very inhomogeneous.
16	Preparation	Plan the study	Statistical analysis plan	You need to have deep understanding of statistics - it	In principle, analysis plans may be acceptable; this	This is an intervention that looks very pertinent for

				seems that you need to have a professional statistician in the team. It is not always the case in research groups (see K. Sijtsma). It will require knowledge to inform researchers and get them acquainted with it	particular intervention may require too much knowledge	drawing reliable conclusions on cause-and-effect questions in research. But it seems to be especially demanding on the knowledge of statistics to be properly utilised.
17	Preparation	Plan the study	Standardization of experimental protocol	Development of shared standards takes time, and may not be applicable in all contexts (e.g. lower income countries)	Development of shared standards takes time, and may not be applicable in all contexts	
18	Preparation	Obtain resources	Researcher trained on study protocol	Training may have limited effect; or require thorough thinking about tests and exams		
19	Data collection	Collect data	Researcher uses standard method for data acquisition	Operational disadvantages can be expected. False confidence and assumption that SOPs are sufficient alone to prevent errors	Good, but depends highly whether uptake can be guaranteed.	standardizing a specific experimental procedure across labs, sites, or operators is USUALLY effective at improving reproducibility
20	Data collection	Collect data	Multi-site testing			
21	Data collection	Store data	Standardized file system structure	The data validation and transformation functionalities will need to be updated for new types of data (Sinaci). ENCORE is based on human readability, which has obvious advantages for the researcher (van Kampen). Still a long way for the community to accept common standards for execution data structures (Kessel)	Acceptability may depend on type of structures and level of commitment	The use of standardized file system structures has been extensively documented in terms of descriptive evidence, and seems promising for use by researchers to increase FAIRness and reproducibility of their data, although quantitative evidence remains limited.

22	Data collection	Store data	Data storage solution	Privacy concerns	Acceptable, but may require funding and training	
23	Data collection	Store data	Direct data deposition			Besides a real-world application of direct data deposition, no studies show evidence for the effectiveness of the intervention. The case study that has been done does give valuable insights into the feasibility and acceptability of this intervention, but the evidence base is limited.
24	Data collection	Document data collection procedures	Researcher uses standardized documentation framework			
25	Data collection	Document data collection procedures	Software tool that documents data retrieval procedures	Very different between the two. See also effort/time/costs/set-up.	Highly dependent on exact intervention	Only two studies; no real comparisons; very different interventions
26	Data collection	Document data collection procedures	Real-time procedure documentation	Very resource intensive and slow, not scalable to single lab research. Even with this much rigour, there were undisclosed protocol deviations. That is logical and proves our thoughts about replications	Mixed, but both skeptics and proponents were included via de consensus panel	Demonstration that a maximal transparency/credibility bundle can be implemented, and that doing so surfaces a null result that survives scrutiny (including a post-hoc audit). Best used as an exemplar of what's POSSIBLE at high cost, not as a low-cost, broadly scalable intervention.

27	Data analysis	Prepare data	Standardized data preparation procedure	Unclear - too heterogeneous		This type of intervention is of low to medium effort and time, under the condition that no new equipment is required, for reasons of standardisation. It seems to be more relevant for purely quantitative fields and some evidence is provided for biomedical-related fields (although only 2 out of the 5 studies provide direct quantitative evidence for the effectiveness), but it is also relevant for Natural sciences and Engineering.
28	Data analysis	Prepare data	Data quality control	Inappropriate use could discard informative samples		
29	Data analysis	Conduct analyses	Researcher uses common method for data analysis	Aggressive standardization of workflows creates a risk of sweeping away true biological variance. Risk of false positives by adopting standard codes as it may sometimes create a false sense of reproducibility if the underlying variables are unmeasured. Time and effort; learning curve;	Researchers are highly motivated to adopt common methods as it lowers publication barriers and helps cross institutional collabs	
30	Data analysis	Conduct analyses	Statistical tool to estimate power			
31	Data analysis	Conduct analyses	Tool to standardize data analysis procedures	well, how useful for our toolbox? It looks way too specific, but maybe we should be that specific. Good to discuss.	Unclear	Both papers are strong, concrete examples of embedding rigor directly into

						a too. But two very different interventions
32	Data analysis	Conduct analyses	Bias analysis method			
33	Data analysis	Validate analysis	Multiverse analysis	Application of different methods of statistical analysis requires time investment.		Application of different methods of statistical analysis seems to effectively increase the reliability of research findings, but it requires significant time investment (more resources and, possibly, create bottlenecks to publication).
34	Data analysis	Validate analysis	Authors use effect-size filter		Is this really an intervention; unclear acceptability.	
35	Data analysis	Document data analysis procedures	Researcher uses standardized documentation framework	Takes time and may feel constraint; then people may not like to use it. High documentation may cause admin fatigue and friction leading to decreased team compliance over long timelines	Barriers to entry is low since these frameworks rely on codebooks and existing repositories rather than heavy software infrastructure	
36	Data analysis	Document data analysis procedures	Open workflow platform	Using cloud computing can lead to dependence on third-party providers. Textflows platform is only available in English		The available studies demonstrate the feasibility of using open workflow platforms, but none shows that using such a platform, compared with not using one, changes reproducibility outcomes

37	Data analysis	Document data analysis procedures	Researchers use tool to document analysis procedures	Well, if I have to trust Claude here as this is a bit above my knowledge: some are very R specific and that is a disadvantage, also dependent on externally hosted services. Low risk of harm, but potential disadvantage can be challenges in adoption for non-technical users.	Standard computing infrastructure without paid software licenses can be used. They align well with FAIR principles and institutional mandates positively impacting acceptability.	Low-cost, low-to-moderate-effort tools with real but unevenly quantified effectiveness, and revisiting the row-fit question (grouping by theme vs. by domain) before finalising
38	Dissemination	Write the report	Use of reporting checklist	The recurring disadvantage across this whole literature is that PASSIVE guideline publication or endorsement alone shows weak, inconsistent effects (2,6,8,10,12) - simply publishing or endorsing a checklist does not reliably change author behaviour. Audit-only studies also cannot cleanly separate the guideline's causal effect from general secular improvements in reporting quality over the same period. The one clear RCT (15) addresses this weakness directly but is a controlled writing exercise with student participants rather than real submitted manuscripts.	Low costs; no additional infrastructure needed. But use depends on internal motivation of researcher	Passive guideline existence/endorsement alone tends to show weak or inconsistent effects on reporting completeness, while active, enforced, or tool-supported interventions could strengthen the overall endorsement
39	Dissemination	Share data	Researchers share data	People may feel reluctance to share data in fear of being critiqued when errors are found. Researcher friction, false sense of security and technical complexities can lead to delays		While making data available is a vital step for reproducibility, these studies show that simply stating data is available, or simply sharing data/code is not enough for

				in publication timelines. Relying on voluntary sharing is risky		reproducibility. While the body of evidence is limited, the study results do point towards the importance of sharing data in a reusable/interoperable way.
40	Dissemination	Share data	Data management tool	SEE NOTES	depends very much on exact intervention	All interventions seem to promote some aspects of openness and transparency. However, there is no - strictly speaking - direct evidence if they promote reproducibility. They do, in principle, but the only evidence is case studies - no comparative studies. Despite this, I believe they worth been included in the toolbox, since they provide some good examples.
41	Dissemination	Peer-review	Peer-reviewer checks reporting	Epecially a dedicated peer reviewer may cost time and effort; but overall it is feasible. Delay in publication timeline, overburdening peer reviewers.	Passive interventions like automated emailing checklists are more acceptable than active ones such as forcing authors to go through rigid submission checks	Might be an effective and feasible method
42	Dissemination	Post-publication	Post-publication data requests			
43	Dissemination	Post-publication	Discrepancy review	legal and professional retaliation and false accusations; resource intense	Low from journals side, higher impact for open science/RI community. A lot of discussion it evokes	Highly acceptable but also fragile considered the operational complexities

44	Dissemination	Post-publication	Replication	This does not improve the quality of the research prospectively.		
----	---------------	------------------	-------------	--	--	--

Under review by the European Commission

Authors: Joeri Tjldink, Mariska Leeftang, Poja Aryal, Ilse de Jong, Vivian Nchanchou Mbanya, Panagiotis Kavouras

Editor: Panagiotis Kavouras