Open Science Indicators: Study Registration

Technical Documentation

This file was initially prepared on 3-4-2024 by Tim Vines and Marcel LaFlamme and is part of the dataset:

Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 7). https://doi.org/10.6084/m9.figshare.21687686.

Named contact Information
Name: Iain Hrynaszkiewicz
ORCID: 0000-0002-9673-5559
Institution: Public Library of Science
Email: ihrynaszkiewicz@plos.org / plos@plos.org

Alternate Contact Information
Name: Lauren Cadwallader
ORCID: 0000-0002-7571-3502
Institution: Public Library of Science
Email: lcadwallader@plos.org / plos@plos.org

This document provides background and technical information for the preliminary release of the study registration indicator. The current documented version is the alpha-release v0.1.

About Study Registrations

Definition

Study registration refers to the plan for a research study, including research questions/ hypotheses, details about the research design, and/or plans for data analysis, which has been made available for public sharing in order to ensure unbiased reporting and support the differentiation of planned and unplanned research directions.

While the term *preregistration* is sometimes used to describe this practice, we follow those who have argued for the more general term *registration*, which acknowledges community variation in the timing of a plan's registration relative to other study milestones.[1]

[1] https://journals.sagepub.com/doi/10.1177/1745691619858427

Selected registries and platforms

31 registries were identified that cover registration of 4 different types of study: clinical trials, systematic reviews, animal studies, and other types of study (or general purpose registries).

Clinical trial registries

The 24 registries in this largest category include the International Clinical Trials Registry Platform (ICTRP, https://www.who.int/clinical-trials-registry-platform) and all of the 20 national and international primary members of the ICTRP. The ICTRP itself is technically not a registry as it only indexes registrations from its network. Furthermore, we have included three secondary registries from Japan which are indexed and registered in the Japan Registry of Clinical Trials (jRCT, https://jrct.niph.go.jp), which is a primary member of the ICTRP. We estimate that these 24 registries (1 ICTRP, 20 primary members, 3 secondary members) contain around 1.8 million registration entries. However, these entries include duplicates as the same clinical trial is often registered and indexed in multiple registries. Given that the ICTRP mirrors every registration in its member registries, we estimate that there are around 900,000 clinical trials overall.

General purpose registries

Open Science Framework (OSF), AsPredicted, and Research Registry are three platforms/registries which are open to study registration submissions from across all fields of research. We estimate that there are 268,000 registrations in these repositories.

Systematic review registries

The International Prospective Register of Systematic Reviews (PROSPERO) and International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) are two registries for systematic reviews. We estimate that there are around 240,000 registered systematic reviews.

Animal study registries

The Animal Study Registry (ASR) and Preclinicaltrials.eu (PCT) are two registries for animal studies and host around 300 registrations in total.

In total, the 31 selected registries host 1.4 million registrations, of which 900,000 are clinical trials and 500,000 are systematic reviews or other study registrations.

Registrations

While the majority of the registrations that we are targeting with the selected registries are for clinical trials, around a third are registrations of non-medical research. This is relevant as the way registrations are used and reported will vary by research community. In the preliminary release of this indicator we will focus on identifying the most common patterns of reporting a registration in a journal article. We can break the act of reporting down into two decisions: *how* does the author report the registration and *where* in the article does it happen?

How are registrations reported within an article?

There are no formalized ways of reporting a registration for a study in an article. The most common approach is to include a brief sentence that states that the trial/study has been registered and provides the registration identifier or a link to the registration, the name or link of the registry, and sometimes a registration date. These sentences are usually part of the main body of the article so it is common to find wide variation in sentence format across journals (and even between versions of articles). Authors also present registration identifiers in inconsistent formats (e.g., NCT12345, NCT 12345, NCT:12345, NCT: 12345).

Where are registrations reported within an article?

Similarly, there are no formal standards around where in the article a registration should be reported, leading to variation both between and within journals. The most common locations are: the last sentence of the abstract, at the beginning of the methods section, in a dedicated trial registration subsection (usually part of methods). Less common are other metadata fields (e.g., data availability statements). Registration protocols can also be provided in the supplementary files.

In general, trial registration is commonly reported multiple times (eg, mentioned at the end of an abstract, within the methods section, and as supplementary material) and each event can look different (abstract events are usually brief, methods events include more details, and in SI the text describing the registration will only appear within the supplemental file itself).

Regular Expressions

The preliminary release of the study registration OSI uses regular expressions to match both registrations and registries to the many different contexts described in the previous section.

It is necessary to distinguish between both registries and registrations as some registries use generic identifiers (e.g., AsPredicted IDs are of the form #12345) and are hence hard to distinguish from other numbers found in the article. In those examples, we rely on the fact that registrations are usually reported in the form of registry-registration pairs such as “The clinical trial was registered at the Dutch Trial Register NTR854”. In this example, the detection of “NTR854” will be considered a registration because the name of its registry (Dutch Trial Register) was found in close proximity.

Identifying registries

For each registry we relied on three types of identifiers to construct regular expressions:

- **Name.** A regular expression that matches the name of the registry, e.g., “Dutch Trial Register”.
- **Abbreviation.** A regular expression to match the abbreviation(s) of a registry, eg, LTR. Some registries have multiple abbreviations because they may have changed names, merged with other platforms, or be a succeeding project.
- **Landing page.** A regular expression to match the landing page of a registry, e.g., https://www.onderzoekmetmensen.nl

Identifying registrations

Identifying registrations requires more complex regular expressions:

- **PID.** The most common and obvious place to start is a regular expression that matches the PID of a registry. For example, the regex “NTR\s{0,1}\d+” will detect NTR854 but also variations with whitespaces between the parts like “NTR 854” or, a tricky case, when a newline makes its way into the identifier when the PID is at the end of a line.
- **DOI.** One registry only references registrations by their DOI and two more register DOIs for their registrations. In all cases, the DOI prefix is unique for these outlets and we can therefore construct regular expressions for each of them. For example, the Animal Study Registry (ASR) has PIDs in the form of “ASR1234567” which are also part of their DOIs: 10.17590/asr.0000131
- **URL.** Authors may also link to registrations on their respective platforms, and these links can include the PID or DOI of the registration but this is not always the case. Therefore, another regular expression will match links to the 31 registries and their landing pages for registrations. For example, https://www.onderzoekmetmensen.nl/en/trial/26529 is the link to the LTR registration NL8960.

Special case: Open Science Framework registrations

The Open Science Framework registrations are identified by their IDs, which are also part of the DOI. For example, an OSF registration with the ID e94t8 has the DOI 10.17605/OSF.IO/E94T8. However, multiple other document types are hosted on OSF, so finding an OSF DOI in the article is not by itself sufficient to determine that the authors are linking to a registration.

To identify OSF registrations we downloaded all 150,000 registrations hosted on OSF, and then checked if any of these OSF IDs or DOIs are mentioned in target articles.

Processing Pipeline

The corpus selected for initial development of the study registration OSI was the Open Science Indicators 2023 Q2 corpus. This was subsequently extended to the full Open Science Indicators corpus from 2019 through 2024 Q1 for the preliminary release.

Processing sections / fulltexts

Each article is provided as an XML file with the content of the article and additional metadata. This XML can come from various sources, so the markup schemas do not necessarily match. However, most of the time, the XML will follow the Journal Article Tag Suite (JATS) standard. Nevertheless, JATS does not fully specify how journals have to implement the standard in all details. Areas of variation might be article section labels and section structure or the implementation of certain additional sections such as a funding statement (e.g., some journals might include it as an additional section in the article while others might use the respective JATS element and still others might implement it in a totally different way). Therefore, one challenge of comparing articles lies in finding and identifying the relevant information from the XML files.

The current OSI workflow is based on matching section titles to manually curated lists of valid terms for a select group of types of sections: methods, data availability statements, funding statements, and abstracts.

For this preliminary release, the processing of these XML files has been revisited to generalize to all section types suggested by the JATS standard[2]. Additionally, customized functions support sections and fields which are not included in the JATS section types such as data availability statements, funding statements, author contributions, or conflicts of interest. The challenge once again here lies in the variety of disciplinary and journal-specific norms around reporting this information.

Detecting Registration Events

With the sections extracted in the previous step, we can now identify registrations in three different locations within each section:

- **Text content.** For each section we extract the text content from the XML and extract potential registration events.
- **External links.** We also check the external links for potential matches with registrations
- **References.** We can globally identify entries in the reference list which point to registrations.

Another location in the article which has not yet been implemented is the detection of registration events described in the text of supplementary materials.

Assembling article-level metrics

All the events that match one of the different regular expressions are potential candidates for registration events. For example, a clinical study might discuss a variety of existing clinical trials and identify those using their ClinicalTrials.gov identifiers. However, these mentions are contextually different from the clinical trial that was registered for the study itself. All of these

[2] https://jats.nlm.nih.gov/publishing/tag-library/1.1/attribute/sec-type.html

individual mentions will be detected and reported by our tool with additional metadata such as the registry name that matched, whether the event was found in the text of the article or in a link, and, if available, section information.

The current implementation of this tool reports all events detected in the main text of the article.

Reference lists

Currently, citations are not processed. Early work on this indicator found that authors almost never cite their DOI for their study registration in the reference list without mentioning it elsewhere in the text.

Supplementary files

Appendices that are included with the article text are treated as sections of the article. All available text elements are concatenated and matched against regular expressions. We do not process the text of supplementary files that are hosted separate from the text of the main article.

Performance

A subset of articles from the 2023 Q2 corpus were manually assessed for registrations. The following table provides an overview of the achieved performance of the preliminary release of the study registration OSI (v0.1). The tool currently performs equally well for both PLOS and PMC articles and achieves an F1 score of 0.94 across both cohorts.

Table 1. Performance scores for the preliminary release of the SR OSI (v0.1)

	Accuracy	F1	Precision	Sensitivity	Specificity
Overall	0.99	0.94	1.00	0.89	1.00
PLOS	0.99	0.93	1.00	0.88	1.00
PMC	0.99	0.95	1.00	0.90	1.00

Looking at the confusion matrices for PLOS and PMC separately we can see that out of the 224 articles considered in the ground truth (GT) set; the tool has only missed two articles that registered trials (one PLOS article and one PMC article).

Confusion Matrices

Table 2. Confusion matrix for the complete GT set containing 224 PLOS and PMC articles.

		GT		
		Yes	No	
OSI:SR	Yes	16	0	*16*
	No	2	206	*208*
		18	*206*	

Table 3. Confusion matrix for the 104 PLOS articles in the GT set.

		GT		
		Yes	No	
OSI:SR	Yes	7	0	*7*
	No	1	96	*97*
		8	*96*	

Table 4. Confusion matrix for the 120 PMC articles in the GT set.

		GT		

		Yes	No	
OSI:SR	Yes	9	0	*9*
	No	1	110	*111*
		10	*110*	