Preregistration

Preregistration: Compliance with Swedish research funders' open access policies

Henrik Danielsson[1], Gustav Nilsonne[2,3], Lovisa Österlund[1], Johanna Nählinder[1]

[1] Linköping University

[2] Stockholm University

[3] Karolinska Institutet

10. July 2019

Study Information

Title	Preregistration: Compliance with Swedish research funders' open access policies
Research questions	The main research question of this project is: how did the introduction of open access publishing mandates by Swedish research funders affect the proportion of open access publication? Additionally, we aim to describe trends in open science publication of Swedish research, with respect to fraction of open access publication and types of open access.

Hypotheses

- H1a: The introduction of open access mandates will be associated with an increase in open access publishing.
- H1b: The introduction of open access mandates will be associated with complete compliance with open access publishing mandates (100%).
- H2: The introduction of open access mandates will be associated with a change in type of open access publishing.

Sampling Plan

Data sources:

- We will collect bibliographic information and funding information from Dimensions or Web of Science. We think that Dimensions has a better coverage than Web of Science, but that will be tested in the beginning of the project. The data source with best coverage of funding information will be used as our primary data source and the other as secondary data source.
- We will collect open access status of published articles from Unpaywall via the R package 'roadoi' (Jahn 2018).

Inclusion criteria:

- Swedish affiliation for the corresponding author. If corresponding author is missing, then affiliation for the first author will be used. For feasibility reasons, we limit the scope to Swedish funding agencies where we know when different open access requirements were introduced. Swedish affiliation is defined by "Sweden" or "Sverige" in the affiliation information for the article as defined above.
- Data will be collected for the following publication years 2008 to 2017. This peridod of time will cover the introduction of open access policies by major funders. The major grant agencies in Sweden who has introduced an open access requirement are the following:

- Vetenskapsrådet - The Swedish Research Council, starting from grants awarded 2010.
- Forte - The Swedish research agency for health, working life and welfare, starting from grants awarded 2012
- Formas - The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning, starting from grants awarded 2010
- Riksbankens Jubileumsfond, starting from grants awarded 2010

Matching procedure:

- The link between articles and funding information is the article doi.
- The funding information is free text in Web of Science (and probably also in Dimensions). We will identify relevant funders with regular expressions. We will search for grant numbers and use them to identify which year the grant was awarded. If no year is identified, we will assume that the grant was awarded 2 years before publication.

Based on the information above, articles with Swedish affiliation will be divided into three groups:

1. Funding information matching the research funders listed above
2. Other funding information
3. No funding information

Existing data

Registration prior to analysis of the data. As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.

Explanation of existing data	All data are available, but we have not performed the matching between data sources. To test the feasibility of the proposed approach, one of the authors searched the first 200 entries from Web of Science to use as pilot data. On the pilot data, two other authors could optimize the regular expression to find different spellings of the grant agencies.
Data collection procedures	See above.
Sample size	The sample size will be all data that is available in the combination of data sources above.
Sample size rationale	The data are limited by when the requirement to publish open access by the main Swedish research funding agencies was introduced. We collect data from 2 years before that was introduced to establish a baseline of degree of open acceess publications. Then we use all data up to 2017 to make sure that we have data for full years. Data is also limited to one data source for each type of data (with the excption of bibliographic data, see text below). We have chosen the data sources with the best coverage of data that we are interested in by doing pilot searches. We will use bibliographic data from Web of Science and Dimensions. Of the data sources that we have evaluated, Web of Science had the best coverage of bibliographic information and funding information that is available for easy download (Google Scholar has better coverage, but no easy way to get the data). However, we have not evaluated the coverage of Dimensions at the time of this pre-registration. The sample of articles is divided into three groups: 1. Funding information matching the research funders listed above 2. Other funding information 3. No funding information The comparison between group 1 and 2 is our main comparison, but group 3 is also included as a comparison. Group 3 could lack funding info for many reasons and is

therefore a weaker comparison than group 2, but it is possible that it is a relatively large proportion of the articles and therefore it is important to include.

Stopping rule No.

Variables

- article publication year
- article open access status
- article funding information
- reprint author affiliation
- article doi
- funding agency startyear of open access publication requirement
- funding agency type of open access publication requirement

Manipulated variables Not applicable.

Measured variables Not applicable.

Indices We will define different types of OA status (Green, Gold, Hybrid, Bronze, and Closed) as rules to use on the Unpaywall data.

Green is defined as "Available from an open repository". The Unpaywall rules for this are is_oa=True AND best_oa_host=repository.

Gold is defined as "Published in an open access journal indexed in DOAJ". The Unpaywall rules for this are is_oa=True AND best_oa_host=publisher AND journal_is_oa=True.

Hybrid is defined as "Published open access in a subscription-based journal under an open license". The Unpaywall rules for this are is_oa=True AND best_oa_host=publisher AND best_oa_license<>""(empty string/null).

Bronze is defined as "Available to read on publisher's site, without a license". The Unpaywall rules for this are is_oa=True AND best_oa_host=publisher AND best_oa_license=""(empty string/null).

Closed is defined as "Not open access". The unpaywall rule for this is is_oa=False.

Design Plan

Study type — **Observational Study**. Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.

Blinding — Data will be collected by two of the authors. The affiliations and oainformation will be given codes. Then two of the other authors will analyze the data without being aware of the meaning of the codes.

Study design — This is an observational longitudinal study with a quasi-experimental design.

Randomization — No.

Analysis Plan

Descriptive analyses — We will tabulate and plot the frequency and proportion of OA articles by funder and year.

Statistical models — We will use an interrupted time series model. OA status will be the outcome. Predictors will be presence of policy (yes/no) and time since introduction of policy. The analysis will be performed as in Hardwicke et al. 2017 (http://rsos.royalsocietypublishing.org/content/5/8/180448, R code available at https://osf.io/y2meu/. Thus, we will fit a general linear model with a logistic link function.

Transformations — No.

Follow-up analyses	Follow-up analyses will be made to investigate if the main result is different depending on affiliation (which university in Sweden).
Inference criteria	Inferences will be based on p-values with alpha <. 05. Two-sided tests will be used so that inferences can be drawn if the introduction of open access mandates is associated with a decrease in open access publishing. Two data sources for bibliographic information will be used, Dimensions and Web of science. The data source with the best coverage (most articles with complete information for our variables) will be used as our primary data source and inferences will be drawn based on results from that data source if the results differ between the data sources.
Data exclusion	No exclusion of data.
Missing data	The degree of missing data from the different data sources will be documented and reported.
Assumptions (optional)	There are several assumptions and limitations for our analysis. We list them here to show that we are aware of them, but we will not conduct any analyses to test the assumptions. • Not all articles are published in journals that is covered in Dimensions or Web of science. On the contrary, there is a rather large difference between academic fields. We are aware of this difference. • Some articles do not have doi:s. The share of articles with doi:s has risen over time. We assume that articles with doi:s follow the same pattern as articles without doi:s. • Some articles do not have funding information. • We assume that articles incorrectly without funding information follow the same pattern as articles with funding information. • Most articles are co-authored. Our study will identify all articles with a Swedish corresponding author. It is plausible that research awarded to a Swedish researcher will in most cases have a Swedish corresponding author

affiliation, but this will not always be the case.

- We assume that the affiliations are correct, i.e. that the words “Sweden” or “Sverige” are correctly used in affiliations.
- OA status of articles can change over time. An article that was not OA when published can become OA later. The other way around is not plausible but this means that there is a bias towards OA in our data. The comparison group will however likely show the same effect.

Exploratory analyses (optional)	No.
Analysis scripts (optional)	TODO

Other

Other (Optional)	No.

References