Meta-Research: Does Scientific Productivity Increase the Publication of Positive Results? Examining Research Groups' Scientific Productivity and Positive Results in a German Clinical Psychology Sample

Louis Schiekiera and Helen Niemeyer
Clinical Psychological Intervention, Freie Universität Berlin

Author Note

Correspondence concerning this article should be addressed to Helen Niemeyer, Clinical Psychological Intervention, Department of Education and Psychology, Freie Universität Berlin, Schloßstraße 1, 12163 Berlin, Email: helen.niemeyer@fu-berlin.de

Abstract

Background: The overrepresentation of positive results in psychology is often attributed in part to publication bias. However, the impact of research group output on the prevalence of positive results has not yet been investigated. The present study examines whether German clinical psychology research groups with high versus low publication outputs differ in the prevalence of positive outcomes in their publications. Methods: Scientific productivity was defined as the ratio of quantitative-empirical publications to the number of academic staff per chair. We analyzed publications authored by clinical psychology researchers at German universities from 2013 to 2022, sourced from PubMed and OpenAlex. After excluding meta-analyses, reviews, and non-empirical studies, 2,280 empirical studies from 99 research groups were identified. We then randomly sampled and coded 300 papers, evenly split between the highest and lowest output quartiles, and examined the first hypothesis. Results: There was no statistically significant difference between the highest and the lowest output quartiles, with both reporting approximately 90% positive results. Higher group paper counts were not associated with more positive results. However, results with partial support were significantly more prevalent in the highest output quartile than in the lowest output quartile. Conclusion: Our results suggest a general excess of positive results in clinical psychology. Contrary to our hypothesis, German clinical psychology research groups with high and low publication outputs do not differ in the prevalence of positive outcomes in their publications.

Keywords: positive results, negative results, scientific productivity, clinical psychology, publication bias, open science, meta-research

Positive results, defined as findings that fully or partially support a tested hypothesis, are typically contrasted with negative results, which do not support (or which contrast with) the hypothesis (Fanelli, 2012). In the quantitative sciences, the criterion for determining whether a hypothesis is supported is often the statistical significance of an effect (Hubbard, 2015). Consequently, positive and negative results are typically distinguished by their statistical significance. Since Fanelli (2010a; 2012) observed that the extent of positive results is particularly pronounced in *psychology and psychiatry* studies compared to other disciplines of science, many studies have investigated positive results in these disciplines. Evidence of high rates of positive results in psychology dates back to the middle of the last century, when Sterling (1959) found that 97% of studies using significance tests in psychology journals reported positive results. Sterling's observations, which he and his colleagues later replicated (Sterling et al., 1995), are in line with more recent studies examining the statistical significance of the first reported hypothesis, which revealed a proportion of positive results ranging from 91-97% in psychology (Fanelli, 2010a, 2012; Scheel et al, 2021; Open Science Collaboration, 2015). However, studies investigating all hypotheses in original studies have found lower rates of positive results, such as 66% (Toth et al., 2021) and 67% (van den Akker et al., 2023).

The high prevalence of positive results has been attributed to several factors, including publication bias (Fanelli, 2012; Scheel et al., 2021), questionable research practices (Scheel et al., 2021), and robust statistical power (Monsarrat & Vergnes, 2018; Sterne et al., 2005), although the latter cannot fully explain the high positive rates in the field of psychology given the often small sample sizes in psychological research (Schäfer & Schwarz, 2019; Szucs & Ioannidis, 2017). Large effects and significant results in underpowered studies suggest that factors such as publication bias and questionable research practices may likewise contribute to these findings (Scheel et al., 2021).

Commonly proposed methods to reduce the overrepresentation of positive results in psychology include registered reports (RRs) and preregistration. Scheel et al. (2021) found that the proportion of positive results in standard psychology reports lay at 96%, compared to 44% in RRs, while van den Akker et al. (2023) did not detect a significant difference between preregistered and non-preregistered studies in terms of the proportion of positive results.

In a study analyzing over 4,600 papers from all scientific disciplines published from 1990 to 2007, Fanelli (2012) observed that negative results are disappearing throughout the sciences, and especially the social sciences. This overrepresentation of positive results has often been framed as a symptom of the transformation of the scientific culture into a 'publish-or-perish' culture, in which increasing competition and pressures to publish distort the cumulative scientific knowledge production (Van Dalen et al., 2012; Tian et al., 2016). In another study, Fanelli (2010b) demonstrated that rates of positive results were correlated with research & development (R&D) expenditures per capita in US federal states, which Fanelli described as indicators of "competitive and productive" academic environments (p. 1). This account suggests that the pressure to publish is primarily determined by macro-level factors (broader political, cultural, and technological[1] processes), but it neglects the fact that pressures to publish might vary in terms of factors on the micro (individual processes) or meso (organizational processes) level as well. From a sociological perspective on science, the *non-publication of negative results* can also be understood as the outcome of an evaluation process (Lamont, 2012; Kjellberg et al., 2013). Such an evaluation process is a complex social interaction involving multiple actors, such as co-authors, and their negotiations (Lamont, 2009; Krüger & Reinhart, 2017), and it is shaped by the social environment, including the organizational context of the respective research group (Friedland & Alford, 1991). While some experimental studies have investigated individual researchers' evaluations of the publishability of study results as a function of their direction (Atkinson et al., 1982; Augusteijn et al., 2023; Chopra et al., 2022; Elson et al., 2020; Epstein, 1990; Mahoney, 1977), metascience studies on rates of positive results and publication bias have overlooked meso-level factors. Therefore, the present study investigates the relevance of a research group's scientific productivity for rates of positive results in clinical psychology. We define scientific productivity as the ratio of the number of quantitative empirical publications within a specific time period to the number of academic staff per research chair.

[1] The creation and development of technical bibliometric infrastructures (e.g., Scholar, Web of Science, etc.) that allow for a more accurate measurement of scientific publication activity is also likely to have increased the pressure to publish.

We propose that positive results might be more prevalent in research groups with a high publication output for two reasons, one *social* and the other *statistical*: First, we assume that on the meso level, the pressure to publish might be more intense in research chairs with high scientific productivity due, for example, to greater expectations of continuous productivity or higher performance orientation, which might in turn result in higher numbers of positive findings as these might be considered to have a higher chance of acceptance for publication (Franco et al., 2014). Second, we suggest that significant results might be more prevalent in research groups with high scientific productivity due to the discussed associations of research funding with sample size (Billingham et al., 2013; Button et al., 2013) and with publication output (Jacob & Lefgren, 2011). In short, research chairs with higher scientific productivity might receive more funding, which may in turn enable larger sample sizes and reduce the possibility of underpowered studies leading to type-II errors (Shreffler & Huecker, 2023).

Present Study

The present study aimed first, to examine the extent to which an excess of positive results, as typically identified in psychology, can also be replicated in clinical psychology and second, to assess the relationship between the proportion of positive results and scientific productivity. We chose to focus on German clinical psychology for two key reasons: First, a comprehensive survey of all published papers in this distinct sample is feasible, allowing us to determine each research chair's publication output over a specific time span. This ability to track the full empirical research output facilitates relative comparisons between different research chairs based on higher or lower publication counts per chair. Second, even though multiple studies have investigated positive results in psychology, it remains unknown whether rates of positive results are likewise exceptionally high in clinical psychology, a sub-discipline that has a strong impact on health systems and society (Doran & Kinchin, 2017; Knapp & Wong, 2020; McDaid et al., 2019).

Our primary hypothesis (H1) was that research groups with lower scientific productivity would report a lower prevalence of positive results than their counterparts with a higher publication output. To test this hypothesis, we collected all empirical studies published by clinical psychology researchers affiliated with German universities between 2013 and 2022. Excluding non-empirical papers, this resulted in a dataset of 2,280 studies. We used a random sample of 300 papers, equally split between the bottom (Q1) and top (Q4) quartiles of scientific productivity, thus taking into account the number of researchers in a research chair, and categorized each paper as reporting positive (full or partial support) or negative results (no support) based on the first reported hypothesis, following the methodology used by Fanelli (2010a; 2012) and Scheel and colleagues (2021).

Furthermore, we tested a secondary hypothesis (H2) by employing logistic regression to explore whether a predictive relationship exists between the publication count of a research group and the likelihood of positive outcomes, i.e. without taking into account the number of researchers in a research chair. This hypothesis was introduced a priori because our measure of scientific productivity might penalize research groups with a higher number of researchers and benefit research groups with a lower number of researchers, depending on whether a high number of early or senior career researchers are affiliated with a respective group.

Methods

Open Practices

Our study was pre-registered on the Open Science Framework (OSF). All data were sourced from published papers. However, since we categorize research chairs according to their scientific productivity, which might be interpreted as a sensitive measure of academic success or failure, we do not publicly disclose our data. Nevertheless, all other aspects of the study methodology, including data collection and analysis procedures, are fully published to ensure transparency and reproducibility of the findings. All R analysis scripts are published on the project's OSF page.

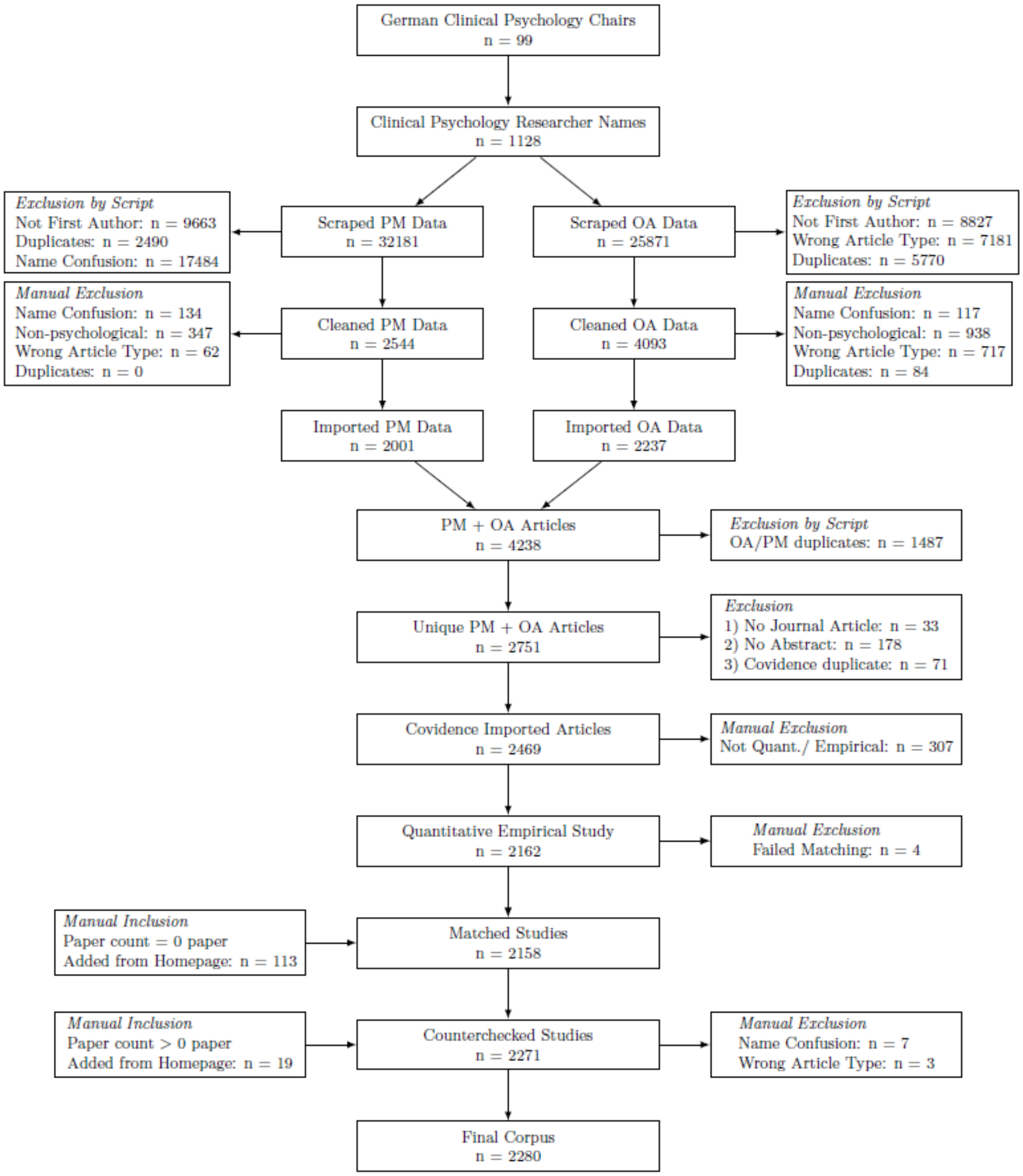
Sample of Publications to Estimate Scientific Productivity

To estimate the scientific productivity of research chairs, we collected all *quantitative-empirical original publications* first-authored by clinical psychology researchers affiliated with German universities from 2013 to 2022. The full procedure is depicted in Figure 1. Exclusions from this collection included meta-analyses, reviews, editorials, comments, corrigenda, errata, letters, and qualitative studies.[2]

[2] The first five data-sampling steps presented are identical to the data collection of the main dataset in Schiekiera et al. (2024), which is presented in the appendix. Since the data collection and definition plays a major role in the computations for the outcome scientific productivity for this article, it is presented in the main text here and in the appendix in Schiekiera et al. (2024).

Figure 1

Data acquisition procedure for the final corpus of all quantitative-empirical original studies first-authored by clinical psychology researchers affiliated with German universities from 2013 to 2022



Note. The final corpus contains abstracts first-authored by German clinical psychology researchers and published between 2013 and 2022; OA = OpenAlex; PM = PubMed.

First, we compiled a list of all state university working groups *(n* = 99) which focused on clinical psychology, psychotherapy, or related fields in Germany, excluding laboratories in hospitals. The chairs are located at 52 different universities and cover all 16 German federal states.

Second, we gathered all names, email addresses, and academic level (predoctoral, postdoctoral, or professorship) of employed researchers from the working group homepages, resulting in a list comprising 1,128 researcher names. Of these, 24 researchers were listed on more than one laboratory or chair web page. We extracted study metadata from (a) PubMed and (b) OpenAlex using the R packages rentrez (Winter, 2017) and openalexR (Priem et al., 2023), respectively. A for loop was used to iterate over the researchers' names *(n* = 1,128) for both OpenAlex and PubMed. The exact search terms for OpenAlex and PubMed can be found in the Appendix. This procedure resulted in a total of 32,181 articles for PubMed and 19,662 articles for OpenAlex.

Third, the extracted OpenAlex and PubMed data were cleaned using a script. Initially, studies authored by researchers who were not on our predefined list were excluded (PM: *n* = 9,663; OpenAlex: *n* = 8,827), along with duplicate articles (PM: *n* = 2490; OpenAlex: *n* = 5,570). Notably, 55% (17,540) of the PubMed articles were attributed to being authored by just five researchers, consisting of two postdocs and three PhD candidates. This anomaly arose because PubMed has a built-in spelling correction function, which mistakenly included researchers with similar names in the corpus. Investigating this discrepancy, one of the present authors (LS) conducted a manual review of the total number of articles authored by these five researchers within our sample, which revealed that only 56 articles were actually first-authored by these researchers. This led to the exclusion of the erroneously included 17,484 entries from our dataset, which we categorized as 'name confusion'. In the case of OpenAlex, additional cleaning steps were performed to exclude data points that did not represent journal articles according to OpenAlex *(n* = 6,209), or which contained the terms 'systematic review', 'meta-analysis', 'comment', 'corrigendum', 'erratum' or 'correction' in the title section (OA: *n* = 972). This step was not necessary for PubMed as these article types could already be excluded through specifications in the search terms.

In the fourth step of our analysis, LS manually screened both OpenAlex and PubMed study and journal titles using an Excel spreadsheet and excluded all non-psychological studies from the corpus. Non-psychological studies were broadly defined as those not reporting the investigation of any association with mental processes, emotions, or behaviors in humans. This criterion led to the removal of 347 articles from the PubMed dataset and 938 articles from the OpenAlex dataset that did not belong to our field of study. For the PubMed dataset, we identified and excluded 134 instances of name confusion (where researchers with similar or common names were incorrectly included) and 62 articles classified under the wrong article types (e.g., review articles instead of original research). The exclusions for the OpenAlex dataset were more extensive: We found 117 instances of name confusion, 717 articles under incorrect article types, and additionally identified and removed 84 duplicates.

In the fifth step, data from OpenAlex and PubMed were merged, with overlapping entries being excluded based on duplicate DOIs *(n* = 1,487). Standardized metadata were extracted using DOIs via the rcrossref library in R (Chamberlain et al., 2022). Crossref indicated that 33 of these entries did not represent journal articles, leading to their exclusion. Subsequently, abstracts were gathered using EndNote 20 (Clarivate, 2023). However, 178 entries lacked identifiable abstracts and were thus removed from the corpus.

The resulting dataset, comprising 2,540 entries, was imported into Covidence (Covidence, 2023), which identified and removed an additional 71 duplicates based on the available metadata. Following this step, the 'Title and Abstract Screening' sample in Covidence consisted of 2,469 entries.

In the fifth step, two researchers (JD, LS) manually annotated a total of *n* = 2,469 studies. During the title and abstract screening phase of our study, *n* = 307 abstracts were excluded for not being quantitative or empirical in nature, resulting in a set of *n* = 2,162 papers.

In the sixth step, we matched these papers with a total of 1,152 authors (including duplications due to associations with more than one research group). An automated matching process using an R-script revealed that the first authors of 4 papers were not listed in our sample and were thus excluded. Additionally, 573 researchers were found to have no quantitative-empirical studies listed after matching.

Consequently, in the seventh step, we counterchecked each researcher without publications listed in our corpus against their respective institutions' homepages. One subject was removed from the dataset because the countercheck revealed they were an institute secretary and not a researcher. For 35 of the 573 researchers without publications, a total of 113 papers were added to the dataset. For the remaining researchers, no publications were identified.

In the eighth step, to further verify the accuracy of our sample, we randomly sampled 101 researchers and checked whether all papers in our dataset were also listed on their publication list web pages. We found that 20 papers had been missed in our initial approach and thus added them to the dataset. Of the 425 papers checked during this step, 7 were excluded due to 'name confusion' and a further 3 were removed as they were posters or review articles.

Our final dataset for the computation of scientific productivity comprised n = 2,280 research papers.

Data Annotation

Papers were annotated utilizing the procedure suggested by Fanelli (2010a; 2012) and Scheel and colleagues (2021) in their studies on rates of positive results: “By examining the abstract and/or full- text, it was determined whether the authors of each paper had concluded to have found a positive (full or partial) or negative (null or negative) support. If more than one hypothesis was being tested, only the first one to appear in the text was considered. We excluded meeting abstracts and papers that either did not test a hypothesis or for which we lacked sufficient information to determine the outcome” (Fanelli, 2010a, p. 8). The criterion of 'insufficient information in the paper' is defined similarly to Scheel et al.'s (2021) study. It includes cases in which the abstract is unclear and the full text of the paper is unavailable for review (Scheel et al., 2021; Fanelli, 2010a). Additionally, this category includes instances in which, even if the full text is accessible, both the abstract and the full text are unclear with respect to the hypotheses tested and/or the conclusions drawn. This implies that each paper in the sample provides a single data point for the primary dependent variable *result type*: whether the first hypothesis was fully supported, partially supported, or not supported at all.

We chose to sample 150 papers each from Quartile 1 (Q1) and Quartile 4 (Q4) of scientific productivity, a numerical index computed as the ratio of the number of quantitative-empirical publications of a research chair (group paper count) to the number of academic staff per chair (number of researchers). This approach was taken to be consistent with the sample sizes used by Fanelli (2010a; 2012) and Scheel and colleagues (2021), in order to estimate the rate of support for the respective first mentioned hypothesis. We randomly sampled a total of 354 papers, of which 54 were excluded from further analysis for the following reasons: 21 were exploratory, 12 were descriptive, 11 lacked a clear hypothesis, 5 were conference abstracts, 2 were not available in full text, 1 was a meta-analysis, 1 had no quantitative hypothesis, and 1 was not an original study. After the first coding round *(n* = 300), additional resampling was necessary: 44 papers for the second round, 8 for the third, and 2 for the fourth. The final sampled dataset comprised 150 papers each for Q1 and Q4. From the included studies, JD and LS annotated the result type: JD rated 45%, LS rated 45%, and both JD and LS rated the remaining 10%. Of the 30 studies rated by both, the raters agreed in 27 cases (*agreement* = 90%; *κ* = .800).

Implementation Period

Data collection was conducted between March and June 2023 and data annotation was completed in February 2024. Data analysis was conducted from March to April 2024.

Statistical Analysis

Variables

The unit of analysis in our study is a journal article, with each article contributing one data point to our dataset. The categories 'full support' and 'partial support' of the dependent variable *result type* are combined into a single category, 'support,' resulting in a binary dependent variable ('support' = 1 vs. 'no support' = 0). The main independent variable is *scientific productivity*, which is defined as the ratio of the number of quantitative-empirical publications of a research chair (*group paper count*) to the number of academic staff per chair (*number of researchers*).

Preregistered Analyses

All analyses were performed using R (R Core Team, 2023). In the present study, we tested two hypotheses:

- **H1**. *Primary hypothesis (scientific productivity):* The rate of positive results in low-output research chairs in German clinical psychology is lower than that in high-output research chairs.
- **H2.** *Secondary hypothesis (per group paper count):* Higher publication counts of research chairs are associated with higher rates of positive results.

For the primary hypothesis (H1), a one-sided proportion test with an alpha level of 5% was conducted to investigate whether the rate of positive results in low-output research chairs (Q1) is lower than that in high-output research chairs (Q4; *scientific productivity*).

Some clinical psychology research groups consist of only one or two persons. Computing the ratio of quantitative-empirical publications to the number of academic staff per chair as an index of scientific productivity might favor these one- or two-person groups and disadvantage larger groups, especially those with a higher number of doctoral students who have only recently joined. Therefore, as part of a sensitivity analysis, we additionally tested whether the rate of positive results (full or partial support) from low-output research chairs in clinical psychology is statistically lower than that in high-output research chairs when excluding (a) all one-person research groups and (b) all one-person and two-person research groups. For the primary hypothesis and the sensitivity analysis, we used the built-in *prop.test* R function to compare rates of positive results between low-output and high-output research chairs (Scheel et al., 2021).

For the secondary hypothesis (H2), we further tested whether higher publication counts of research chairs (*group paper count*) are associated with higher rates of positive results, using logistic regression with an alpha level of 5%. The outcome was binary ('support' = 1 vs. 'no support' = 0) and the predictor was metric (no. of papers). Thus, we hypothesized a positive relationship between outcome and predictor. Fanelli (2012) used this method to predict the result as a function of publication year. For this analysis, we used the built-in *glm* R function for logistic regression.

Exploratory Analyses

In addition to testing the two hypotheses, we conducted the following exploratory analyses to provide a broader context for our findings: First, we compared the rates of (a) full support ('full support' = 1 vs. 'no support OR partial support' = 0) and (b) partial support (partial support' = 1 vs. 'no support OR full support' = 0) between low-output and high-output research chairs to investigate whether rates of positive results differ as a function of scientific productivity when using different operationalizations of positive results than the categorization 'full support OR partial support' = 1 vs. 'no support' = 0 suggested in the literature (Fanelli, 2010a; Scheel et al., 2021). Second, we also compared the rates of positive results in our sample with those reported by Fanelli (2010a) and Scheel et al. (2021) in order to place our findings in the context of previous research. Third, we extended our logistic regression model by including additional variables such as the individual paper count of researchers, publication year, and the number of researchers in a group in order to explore their potential associations with rates of positive results. Lastly, we analyzed reporting patterns between the top and bottom quartiles of scientific productivity, focusing on the use of the term "significant" and the explicit mention of hypotheses.

Results

Descriptive Statistics

Researchers

The qualification levels among the *n* = 1,151 researchers in our study were distributed as follows: The majority (712, 61.86%) were predoctoral researchers and PhD students, followed by postdocs (321, 27.89%), and professors (118, 10.25%). The mean number of first-authored quantitative-empirical original works published per researcher between 2013 and 2022 lay at 2.07 (*SD* = 3.70). Mean first authorship counts differed by qualification level: Predoctoral researchers and PhD students had a mean of 0.61 (*SD* = 1.12) of first-authored papers, postdocs a mean of 4.01 (*SD* = 4.61), and professors a mean of 5.60 (*SD* = 5.75).

Research groups

On average, research groups consisted of 11.63 researchers per group (*SD* = 9.15, *Median* = 10, *Range* = [1, 55]). The mean number of first authorships of quantitative-empirical original works per group published between 2013 and 2022 lay at 24.03 (*SD* = 21.59, *Median* = 18, *Range* = [0, 153]). Across the groups, the mean scientific productivity (defined as the ratio of quantitative-empirical publications to the number of academic staff per chair) was 2.36 (SD = 1.79, Median = 1.89, Range = [0, 10.2]). To compare the bottom quartile (Q1) and the top quartile (Q4) of scientific productivity from the *n* = 99 research chairs, we assigned 25 research chairs to the bottom quartile and 24 research chairs to the top quartile, and 25 each to Q2 and Q3. Table 1 and Figure 2 show the distribution of number of publications and number of researchers as well as the different quartiles of scientific productivity of the research chairs.

Full publication dataset

The publication years of the 2,280 articles in our full dataset ranged from 2013 to 2022, with a mean publication year of 2018.29 (*SD* = 2.99). 2,183 (95.75%) were written in English, while 97 (4.25%) were written in German. The most frequent journals in the corpus, with their respective counts, were Frontiers in Psychology *(n* = 86), PLOS ONE *(n* = 80), Psychiatry Research *(n* = 54), Behaviour Research and Therapy *(n* = 52), and Journal of Behavior Therapy and Experimental Psychiatry *(n* = 49). In the full publication dataset, the distribution of scientific productivity quartiles was as follows: Researchers from Q1 accounted for 13.29% of all publications *(n* = 303), researchers from Q2 for 22.11% *(n* = 504), researchers from Q3 for 34.12% *(n* = 778), and researchers from Q4 for 30.48% *(n* = 695).

Table 1

Descriptive statistics for the four quartiles of scientific productivity for n = 99 German clinical psychology research chairs

Quartile	Metric	*M*	*SD*	*Median*	*Range*
Q1	Scientific Productivity	0.90	0.41	1.00	[0, 1.25]
	Group Paper Count	12.40	14.70	7.00	[0, 65]
	Number of Researchers	11.96	12.38	8.00	[1, 55]
Q2	Scientific Productivity	1.62	0.18	1.67	[1.29, 1.89]
	Group Paper Count	21.28	12.87	17.00	[5, 57]
	Number of Researchers	13.24	8.06	12.00	[3, 34]
Q3	Scientific Productivity	2.30	0.25	2.27	[2, 2.67]
	Group Paper Count	31.56	19.32	27.00	[4, 66]
	Number of Researchers	13.64	8.22	11.00	[2, 31]
Q4	Scientific Productivity	4.73	2.16	3.94	[2.79, 10.20]
	Group Paper Count	31.17	30.47	29.00	[5, 153]
	Number of Researchers	7.50	5.79	7.50	[1, 24]

Note. Group Paper Count = Number of publications per group between 2013 and 2022.

Hypotheses

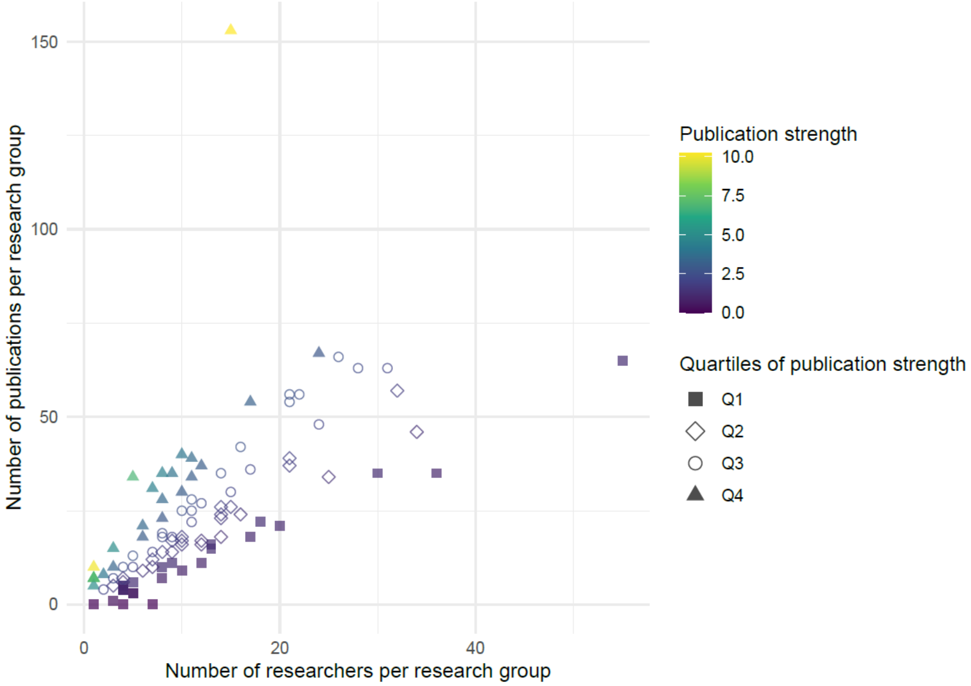
H1: Primary Hypothesis – Scientific Productivity

Among the sampled papers, 135 out of 150 from Q1 (low scientific productivity) and 134 out of 150 from Q4 (high scientific productivity) reported full or partial support for the first mentioned hypothesis. This equals a positive result rate of 90.00% for Q1 papers (95% confidence interval (95% *CI* = [84.04, 94.29]) and 89.33% for Q4 (95% *CI* = [83.25, 93.78]; see Fig. 3). The observed difference of −0.67% was not statistically significant according to the preregistered one-sided proportions test ($\chi^2(1) = 0.00$, $p = .500$). The overall rate of positive results was 89.67%.

Consequently, the hypothesis that the rate of positive results in the top quartile is lower than that in the bottom quartile (H1) was rejected.

Figure 2

Number of publications and number of researchers per research group and scientific productivity for n = 99 German clinical psychology research chairs



Note. Number of publications per group between 2013 and 2022.

Sensitivity analyses revealed that even after excluding all papers from research chairs consisting solely of one person *(n* = 5 papers from Q4, none from Q1) and all one- and two-person research chairs *(n* = 6 papers from Q4, none from Q1), the observed differences of -1.03% and -1.11% were not statistically significant according to two preregistered one-sided sensitivity proportions tests (no one-person research groups: χ2(1) = 0.01, *p* = .540; no one- and two-person research groups: χ2(1) = 0.01, *p* = .548).

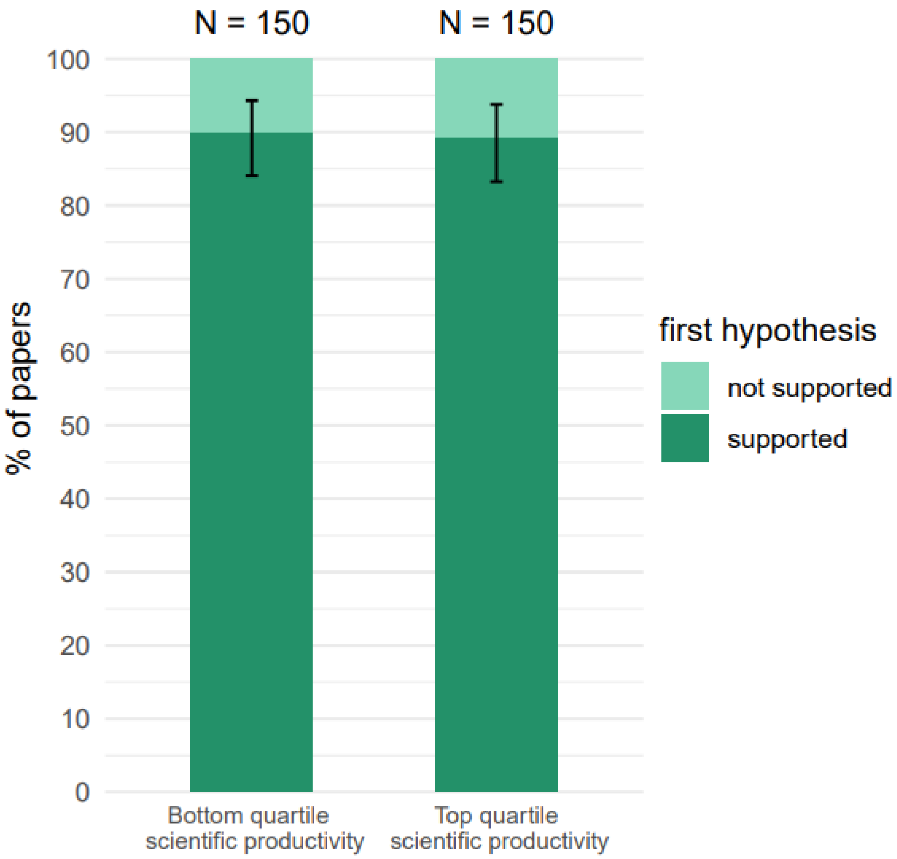
H2: Secondary Hypothesis – Group Paper Count

The logistic regression model testing the association between higher publication counts of research chairs and higher rates of positive results showed very weak explanatory power (*Tjur's* R^2 = .00). The effect of group paper count on positive results was not statistically significant (beta

= 0.00, 95% *CI* [-0.00, 0.02], p = .356). Therefore, the hypothesis that higher publication counts of research chairs are associated with higher rates of positive results (H2) was rejected.

Figure 3

Rates of positive results for high-output and low-output research chairs. Error bars show 95% confidence intervals around the observed rate of positive results.



Exploratory results

Full and partial support

Ninety-four out of 150 (62.67%) papers from Q1 and 78 out of 150 (52.00%) papers from Q4 reported *full* support for the first mentioned hypothesis. In contrast, 41 out of 150 (27.33%) papers from Q1 and 56 out of 150 (37.33%) papers from Q4 reported *partial* support for the first mentioned hypothesis. The observed differences of −10.67% for full support and +10% for partial support were statistically significant according to two non-preregistered one-sided proportions tests (full support: $\chi^2(1) = 3.07$, p = .040; partial support: $\chi^2(1) = 2.99$, p = .042).

Rates of positive results in psychology papers in other studies using Fanelli's method

We compared the rate of positive results found in our analyses with those reported by Fanelli (2010a) and Scheel et al. (2021), adopting their approaches to estimate the rate of positive results in psychology and psychiatry papers. The difference in rates of positive results between our sample and those reported by Fanelli (89.67% − 91.49% = -1.82%) was not statistically significant ($\chi^2(1) = 1.84$, $p = .667$). However, the difference between our sample and the standard psychology papers[3] (excluding registered reports) examined by Scheel and colleagues (89.67% − 96.05% = -6.38%) was statistically significant ($\chi^2(1) = 4.66$, $p = .031$).

Logistic regression associations

Furthermore, we introduced three further variables into our logistic regression model from H2: the *paper count of an individual*, the *publication year*, and the *number of researchers* in a group. The logistic model likewise showed very weak explanatory power here (*Tjur's* $R^2 = .01$). All regression coefficients were not statistically significant (*group paper count*: beta = 0.00, 95% *CI* [-0.00, 0.02], $p = .313$; *individual paper count*: beta = 0.00, 95% *CI* [-0.06, 0.05], $p = .702$; *publication year*: beta = -0.11, *95% CI* [-0.24, 0.03], $p = .125$; *number or researchers*: beta = 0.00, 95% *CI* [-0.03,0.03], $p = .884$).

Reporting differences between Q1 and Q4

Of all $n = 288$ papers written in English ($n = 12$ in German), 38.00% in Q1 and 26.67% in Q4 used the pattern "significant*" when describing the results of the first mentioned hypothesis; this difference between the two quartiles was statistically significant ($\chi^2(1) = 3.90$, $p = .048$). Furthermore, 36.67% of English-language papers in Q1 explicitly mentioned the hypothesis (pattern “hypothes*"), compared to 39.33% in Q4; this difference was not statistically significant ($\chi^2(1) = 0.13$, $p = .721$). A common way of introducing hypotheses was through the textual pattern "hypothesize/hypothesise," found in $n = 88$ (30.56%) of all papers written in English. Only $n = 3$ (1.04%) of all papers written in English used the pattern "test* the hypothes*", which Fanelli (2010) and Scheel et al. (2021) used for their sampling of standard quantitative studies in psychology.

[3] Scheel and colleagues (2021) use the term "standard psychology papers" to contrast them with registered reports, meaning that standard papers are non-registered report studies.

Discussion

In the present study, we assessed the percentage of clinical psychology articles that supported the first hypothesis and found that overall, the proportion of positive results was high, with approximately 9 in 10 studies (89.67%) reporting positive results. Contrary to our primary hypothesis, we did not find a notable discrepancy between research chairs with high and low research output (89.33% vs. 90.00%). This suggests that the high rate of positive results is pervasive, and is not significantly influenced by the level of research productivity.

Our findings differ from those of Fanelli (2010b), who reported that pressures to publish, measured at the federal state level in the US, increased scientists' bias toward positive results. Using the scientific productivity of research groups as a meso-level proxy of academic productivity, we did not confirm this pattern. Further analyses revealed that there were also no significant differences after excluding all publications from one-person chairs and from one- and two-person chairs. Moreover, higher publication counts of research chairs were not associated with higher rates of positive results in a logistic regression model. Our preregistered analyses suggest that the high proportion of positive results in clinical psychology cannot be differentially explained by scientific productivity or group paper count.

However, our exploratory analyses suggested differences between high and low research outputs in terms of the extent of reporting full support and partial support for the first reported hypothesis: Surprisingly, publications from research chairs with low research output more frequently reported full support than did publications from research chairs with high research output. Conversely, publications from research chairs with low research output less frequently reported partial support compared to publications from research chairs with high research output. This pattern might be partly explained by differences in reporting style, with the bottom quartile mentioning the pattern "significant*" more often than the top quartile. Contrary to our expectation, researchers in groups with low scientific productivity more often presented their findings as fully rather than partially positive. We can only speculate about the reasons for this. According to expected utility theory, individuals with higher resources are less attracted by prospective gains but also less discouraged by possible loss compared to individuals with lower resources (Mongin, 1997; Passarelli & Del Ponte, 2020). Thus, researchers in groups with low productivity may perceive the non-publication of a paper due to negative results as a greater loss compared to researchers in highly productive groups, who already have a steady stream of publications and might therefore be more willing to report partially positive, mixed, or negative results.

Furthermore, the proportion of positive results in our sample of clinical psychology studies (89.67%) was significantly lower than the proportion of 96.05% reported by Scheel and colleagues (2021) but did not differ significantly from the proportion of 91.5% reported by Fanelli (2010a). It is challenging to find a clear explanation for why our observations of positive results are similar to those of Fanelli (2010a) but lower than those of Scheel and colleagues (2021), because the latter study represents a conceptual replication of the former, with everything held constant except for the publication year. Scheel (2021) analyzed a more recent dataset (2013 to 2018), while Fanelli (2010a) did not place any restriction on year of publication. Neither study focused exclusively on clinical psychology.

Strengths and limitations

A strength of the current study is that we included all empirical and quantitative original studies from the clinical psychology sample, and not merely those that employed the phrase 'test* the hypothes*' (Fanelli, 2010a; Scheel et al., 2021). By not restricting our dataset to studies that used specific phrasing to describe their hypothesis-testing, we minimized the potential bias that might be introduced by selectively sampling only those studies explicitly mentioning this pattern, since only *n* = 3 (1.04%) of all papers written in English used the pattern "test* the hypothes*" in our sample. However, the inclusion of all papers comes with challenges in terms of coding hypotheses, as determining what constitutes a hypothesis can be difficult when the language used is ambiguous.

While the present study investigated the rates of positive results in the top and bottom quartile of scientific productivity, we did not analyze those in the second and third quartile of this variable. This approach may limit our understanding of the overall distribution and variability of positive results across the spectrum of scientific productivity. By focusing on the extremes, we may have missed patterns that might occur in the middle quartiles, which could provide a more nuanced understanding of the factors influencing positive results in clinical psychology.

Our research focused exclusively on German clinical psychology, limiting the ability to generalize our findings to other geographical areas and different fields of psychology or related disciplines. This specificity may obscure broader trends, rendering it challenging to determine whether the observed patterns are unique to German research culture, clinical psychology as a whole, or specifically to German clinical psychology. It would not have been feasible to additionally measure the relative scientific productivity of clinical psychology in other countries as comprehensively as for German clinical psychology, given the extensive data collection, cleaning and coding steps in this analysis.

The operational definitions of 'positive results' and 'scientific productivity' might have influenced the interpretation of the data. Alternative definitions of 'positive results' could include the investigation of all hypotheses in a study (van den Akker et al., 2023) or the distinction between solely positive results and mixed or negative results (Schiekiera et al., 2024). Moreover, our definition of scientific productivity may disproportionately benefit research chairs with a small number of researchers while penalizing those with a larger number of researchers. To mitigate this, however, we also controlled for the raw group and individual paper count.

Conclusion

This study examined the proportion of positive results within research groups in German clinical psychology, analyzing the association between scientific productivity and positive outcomes. Contrary to our hypothesis, the degree of scientific productivity did not differentially explain high rates of positive outcomes. Furthermore, our study showed that partially supported hypotheses were more prevalent in the highest output quartile than in the lowest, a difference that may be influenced by variations in reporting styles between different scientific productivity levels. These findings underscore the need for further research to explore the complex dynamics of evaluative processes that lead to the non-publication of negative results, considering factors at the macro, meso, and micro levels.

References

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship?. Journal of Counseling Psychology, 29(2), 189.

Augusteijn, H. E., Wicherts, J., Sijtsma, K., & van Assen, M. A. (2023). Quality assessment of scientific manuscripts in peer review and education.

Billingham, S. A., Whitehead, A. L., & Julious, S. A. (2013). An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. BMC medical research methodology, 13, 1-6.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews neuroscience, 14(5), 365-376.

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2022). Rcrossref: Client for various crossref apis [R package version 1.2.0]. https://CRAN.R-project.org/ package=rcrossref

Chopra, F., Haaland, I., Roth, C., & Stegmann, A. (2022). The null result penalty. CESifo Working Paper No. 9776.

Clarivate. (2024). EndNote 20. www.support.clarivate.com/Endnote/

Covidence. (2023). Covidence review software. www.covidence.org.

Doran, C. M., & Kinchin, I. (2017). A review of the economic impact of mental illness. Australian Health Review, 43(1), 43-48.

Elson, M., Huff, M., & Utz, S. (2020). Metascience on peer review: Testing the effects of a study's originality and statistical significance in a field experiment. Advances in Methods and Practices in Psychological Science, 3(1), 53-65.

Epstein, W. M. (1990). Confirmational response bias among social work journals. Science, Technology, & Human Values, 15(1), 9-38.

Fanelli, D. (2010a). "Positive" results increase down the hierarchy of the sciences. PLoS ONE, 5(4), e10068.https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2010b). Do pressures to publish increase scientists' bias? An empirical support from US States Data. PloS one, 5(4), e10271.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics, 90(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502-1505. 10.1126/science.1255484

Friedland, R., & Alford, R. (1991). Bringing Society Back In: Symbols, Practices and Institutional Contradictions. In W. Powell & P. Dimaggio (Hrsg.), The New Institutionalism in Organizational Analysis (S. 232–263). University Of Chicago Press.

Hubbard, R. (2015). Corrupt research: The case for reconceptualizing empirical management and social science. Sage Publications.

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. Journal of public economics, 95(9-10), 1168-1177.

Kjellberg, H., & Mallard, A. (2013). Valuation Studies? Our Collective Two Cents. Valuation Studies, 1(1), 11–30. https://doi.org/10.3384/vs.2001-5992.131111

Knapp, M., & Wong, G. (2020). Economics and mental health: the current scenario. World Psychiatry, 19(1), 3-14.

Krüger, A. K., & Reinhart, M. (2017). Theories of Valuation—Building Blocks for Conceptualizing Valuation Between Practice and Structure. Historical Social Research.

Lamont, M. (2009). How professors think: Inside the curious world of academic judgment. Harvard University Press.

Lamont, M. (2012). Toward a Comparative Sociology of Valuation and Evaluation. Annual Review of Sociology, 38(1), 201–221. https://doi.org/10.1146/annurev-soc-070308-120022

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive therapy and research, 1, 161-175.

McDaid, D., Park, A. L., & Wahlbeck, K. (2019). The economic case for the prevention of mental illness. Annual review of public health, 40, 373-389.

Monsarrat, P., & Vergnes, J.-N. (2018). The intriguing evolution of effect sizes in biomedical research over time: Smaller but more often statistically significant. GigaScience, 7(1), gix121.

Mongin, Philippe. (1997). Expected utility theory. Handbook of Economic Methodology. 342-350.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Passarelli, F., & Del Ponte, A. (2020). Prospect theory, loss aversion, and political behavior. In Oxford Research Encyclopedia of Politics.

Preregistration for Qualitative Research Template. (2018). Retrieved from https://osf.io/b6xmd/

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in psychology, 10, 442717.

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. Advances in Methods and Practices in Psychological Science, 4(2), 25152459211007467.

Schiekiera, L., Diederichs, J., & Niemeyer, H. (in press). Classifying positive results in clinical psychology using natural language processing. *Zeitschrift für Psychologie: Topical Issue Natural Language Processing in Psychology*. Hogrefe.

Shreffler, J., & Huecker, M. R. (2023). Type I and Type II errors and statistical power. In StatPearls. StatPearls Publishing. PMID: 32491462.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. Journal of the American Statistical Association, 54(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. The american statistician, 49(1), 108–112.

Sterne, J. A., Becker, B. J., & Egger, M. (2005). The funnel plot. Publication bias in meta-analysis: Prevention, assessment and adjustments, 73-98.

Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. Front Hum Neurosci, 11, 390. https://doi.org/10.3389/fnhum.2017.00390

Tian, M., Su, Y., & Ru, X. (2016). Perish or publish in China: Pressures on young Chinese scholars to publish in internationally indexed journals. Publications, 4(2), 9.

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. Journal of Business and Psychology, 36, 553–571.

van Dalen, H. P., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. Journal of the American Society for Information Science and Technology, 63(7), 1282-1293.

van den Akker, O. R., van Assen, M. A., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2023). Preregistration in practice: A comparison of preregistered and nonpreregistered studies in psychology. https://osf.io/preprints/metaarxiv/fhdbs/

Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API (No. e3179v2). PeerJ Preprints.

Author Contributions:

- Conceptualization: L.J. Schiekiera, H. Niemeyer
- Methodology: L.J. Schiekiera, H. Niemeyer
- Software: L.J. Schiekiera
- Validation: L.J. Schiekiera
- Formal Analysis: L.J. Schiekiera
- Investigation: L.J. Schiekiera, H. Niemeyer
- Resources: H. Niemeyer
- Data Curation: L.J. Schiekiera
- Writing – Original Draft: L.J. Schiekiera, H. Niemeyer
- Writing – Review & Editing: L.J. Schiekiera, H. Niemeyer
- Visualization: L.J. Schiekiera,
- Supervision: H. Niemeyer
- Project Administration: L.J. Schiekiera, H. Niemeyer
- Funding Acquisition: H. Niemeyer

Conflicts of Interest: The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgments: Thanks to the Berlin University Alliance for funding this project. Special thanks to Kristina Eichel and Jonathan Diederichs for their help in this research project. Special thanks as well to Sarah Mannion for correcting the manuscript.

Funding: The project was funded by the Berlin University Alliance (312_OpenCall_3).

Appendix

Search terms for the publication sample.

OpenAlex: Scraping abstracts using the openalexR library

```
library(openalexR)
oa_fetch(entity = "authors",
   display_name = name)
   oa_fetch(
   entity = "works",
   author.id = gsub(pattern = "https://openalex.org/",
   replacement = "",
   name),
   from_publication_date = "2013-01-01",
   to_publication_date = "2022-12-31",
   verbose = TRUE
)
```

PubMed: Scraping abstracts using the rentrez library

```
library(rentrez)
entrez_search(
db = "pubmed",
api_key = api_key,
tool = tool,
email = email,
term = paste0("(", name, ’("NAME"[Author]) AND ("2013"[Date - Publication]:
"2022"[Date - Publication]) NOT (Review[Publication Type]) NOT (Systematic
Review[Publication Type]) NOT (Meta-Analysis[Publication Type]) NOT (Case
Reports[Publication Type]) NOT (Editorial[Publication Type]) NOT
(Letter[Publication Type]) NOT (Editorial[Publication Type]) NOT
(News[Publication Type])’)
)
```