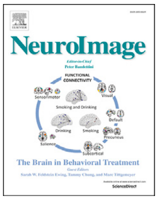
Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals

Denes Szucs [a,*], John PA. Ioannidis [b]

[a] University of Cambridge, Department of Psychology, UK

[b] Meta-Research Innovation Center at Stanford (METRICS) and Department of Medicine, Department of Epidemiology and Population Health, Department of Biomedical Data Sciences, And Department of Statistics, Stanford University, Stanford, CA, USA

ABSTRACT

We evaluated 1038 of the most cited structural and functional (fMRI) magnetic resonance brain imaging papers (1161 studies) published during 1990–2012 and 270 papers (300 studies) published in top neuroimaging journals in 2017 and 2018. 96% of highly cited experimental fMRI studies had a single group of participants and these studies had median sample size of 12, highly cited clinical fMRI studies (with patient participants) had median sample size of 14.5, and clinical structural MRI studies had median sample size of 50. The sample size of highly cited experimental fMRI studies increased at a rate of 0.74 participant/year and this rate of increase was commensurate with the median sample sizes of neuroimaging studies published in top neuroimaging journals in 2017 (23 participants) and 2018 (24 participants). Only 4 of 131 papers in 2017 and 5 of 142 papers in 2018 had pre-study power calculations, most for single t-tests and correlations. Only 14% of highly cited papers reported the number of excluded participants whereas 49% of papers with their own data in 2017 and 2018 reported excluded participants. Publishers and funders should require pre-study power calculations necessitating the specification of effect sizes. The field should agree on universally required reporting standards. Reporting formats should be standardized so that crucial study parameters could be identified unequivocally.

1. Introduction

The number of participants is in general low in cognitive neuroscience and neuroimaging. Thus, it has been pointed out that statistical power that depends on sample size is also likely to be low in these studies. Consequently, many false negative outcomes, imprecise measurements, exaggerated published statistically significant effect sizes and high false report probability can be expected in this field (Desmond and Glover, 2002; Murphy and Garavan, 2004; Yarkoni, 2009; Ingre, 2013; Lindquist et al., 2013; Ioannidis, 2008, 2005a,b; Button et al., 2013; Poldrack et al., 2017; Szucs and Ioannidis, 2017a,b; Turner et al., 2018; Geuter et al., 2018; Cremers et al., 2017; Petersson et al., 1999; Zandbelt et al., 2008).

In the null hypothesis significance testing (NHST) framework statistical power is defined as the probability of getting a statistically significant test result (rejecting the null hypothesis) *given* that a well-defined alternative hypothesis with a specified effect size is true and therefore the null hypothesis is false (for extended review see Szucs and Ioannidis, 2017a). It is often thought that statistical power is only important for studies because low power precludes the detection of existing effects. However, studies with low power also have other serious problems: First, low power increases false report probability, the probability that statistically significant findings are in fact false (Ioannidis, 2005; Szucs and Ioannidis, 2017b). Second, using low sample sizes (and therefore having low power) leads to noisy measurements due to high sampling variability. Hence, many studies with low power will likely report widely different results. Third, if mostly only small sample size NHST studies with statistically significant results are published, then these will inevitably report exaggerated (large) effect sizes even if the true phenomenon produces small effect sizes (Yarkoni, 2009; Geuter et al., 2018). This is so because by using small sample sizes and therefore small degrees of freedom only relatively large effects have a chance to pass traditional statistical significance testing thresholds (e.g. $\alpha = 0.05$). Such large effects may occur occasionally due to sampling variability. Large effects can also be the result of p-hacking when analytical manipulation makes the results from these small studies to pass the significance threshold. Many such exaggerated published effects from small studies will then distort the literature and may also be picked up by meta-analyses, thus further resulting in exaggerated meta-analytic effect sizes.

In the NHST framework the larger is the sample size, the to-be-detected effect size and the α level the larger is statistical power. Increasing the α level (e.g. from 0.05 to 0.10) is problematic in neuroimaging because of the multiple testing problem (Yarkoni, 2009).

[*] Corresponding author.
E-mail address: ds377@cam.ac.uk (D. Szucs).

https://doi.org/10.1016/j.neuroimage.2020.117164
Received 17 October 2019; Received in revised form 7 July 2020; Accepted 9 July 2020
Available online 15 July 2020

Further, increasing the α level will also greatly increase the amount of false positive results if the null hypothesis is true (for discussions regarding the α level see Benjamin et al., 2018 and Lakens et al., 2018; Wasserstein et al., 2019; Amrhein et al., 2019; McShane et al., 2019). According to the NHST framework it is crucial to determine statistical power so that optimal decisions could be made regarding rejecting or not rejecting the null hypothesis (Neyman and Pearson, 1933; for review see Szucs and Ioannidis, 2017a). However, in neuroimaging the effect sizes sought are often difficult to determine. This is partly due to the fact that the actually measured signal changes are most often not communicated, statistically significant effect size reports are likely to be exaggerated and studies often only aim to reject a hypothesis of zero effect rather than a well-justified numerically expressed effect (Szucs and Ioannidis, 2017a; b). Due to uncertainty about effect sizes, many investigators have suggested that the most straightforward way to increase power in neuroimaging would be through a parameter researchers have some control of: by using sample sizes justified by power calculations based on realistic expected effet sizes (Yarkoni, 2009; Desmond and Glover, 2012; Geuter et al., 2018; Turner et al., 2018; Suckling et al., 2014; Poldrack et al., 2017). Increasing sample sizes would also increase measurement precision. However, in contrast to their theoretical and practical importance it is rare to see power calculations in papers published in many disciplines.

Here, we aimed to extend previous work on scrutinizing sample sizes in neuroimaging and additionally, examined the prevalence of power calculations. First, we determined participant numbers in the most cited experimental functional magnetic resonance imaging (fMRI) studies published between 1990 and 2012. We were especially interested in highly cited studies because (by definition) they are very influential in the scientific literature and because they are likely to set standards for many researchers. Previously (Szucs and Ioannidis, 2017a) we observed that, on average, more participants were examined in papers in medically oriented than in cognitive neuroscience journals. So, for comparison we also report participant numbers in the most cited structural MRI (sMRI) and fMRI clinical studies that examined patients. Further, in order to monitor progress in participant numbers we also determined sample sizes in studies published in 4 top neuro-imaging journals in 2017 and 2018 (the latest available complete years before we did data collection). Moreover, to learn whether participant numbers were set in a principled way in recent papers we have collected data about the frequency and method of (pre-study) power calculations in studies published in 2017 and 2018.

2. Methods

2.1. Highly cited papers: Identification and data extraction

We evaluated sample size data from 1038 of the most cited sMRI and fMRI papers (1161 studies) published during 1990–2012 and 273 papers (302 studies) published in 4 top neuroimaging journals during 2017 and 2018. By 'paper' we mean a publication unit published as a formal paper in a journal, whereas by 'study' we mean the individual studies reported in papers. Some papers reported more than one MRI study. Hence, the number of studies is higher than the number of papers. We evaluated only fMRI studies whereas some papers also included purely behavioral, electro-encephalography, and other types of non-eligible studies.

First, we queried the Scopus (scopus.com) search engine for the 1500 most highly cited 'articles' using magnetic resonance imaging (MRI) published from 1990 onwards in the 'neuroscience' field. The date of query was May 25, 2017 and it returned papers published between 1990 and 2012. The search term was TITLE-ABS-KEY (*MRI*) AND DOCTYPE (ar) AND PUBYEAR > 1989 AND (LIMIT-TO (SUBJAREA, "NEUR")). The query (see main text) generated a comma separated text file. During the process of data extraction we added additional records to this file describing participant numbers and study types.

We aimed to examine primary empirical research reports that used in vivo sMRI or fMRI to study brain structure and function in humans. So, we excluded misclassified review papers, methodological papers, meta-analyses of published findings, post-mortem studies, case studies, animal studies, behavioral papers, theoretical papers, modelling papers, papers on surgery which only used MRI to aid surgery, non-brain MRI papers (e.g. MRI of the chest and muscles), and papers with other than MRI technology (positron emission tomography, electro-encephalography, computed tomography).

Specifically, we first read titles and abstracts queried from the Scopus database. For all studies of interest we accessed full text pdf files where possible and we confirmed whether a certain paper was appropriate for study. If a paper was appropriate for study then we manually extracted participant numbers from most papers by reading the 'Participants', or equivalent, sections of full text pdf files. In case of uncertainty about participant numbers other sections of papers were also examined. We could not access pdf files for 48 relevant papers but we were able to extract participant information from abstracts and online full texts. We could not access participant data for 9 relevant papers, so they were not considered for analysis (marked as type = 'x … ' in the data file).

In remaining sample there were 1098 papers (1223 studies). The journals most represented in our sample are shown in Supplementary Table 1. These studies could be sorted into 6 major categories:

(1) Experimental fMRI cognitive neuroscience studies with normal adults (experimental studies). The primary concern of these studies was the understanding of brain structure and function and they did not have primary clinical relevance. Most of the experimental fMRI studies compared brain activity across two or more experimental conditions in a single group of participants. The approximate topics of experimental fMRI papers are shown in Supplementary Table 2.

(2) Cognitive neuroscience sMRI studies typically used structural data to support the interpretation of fMRI data; to gain anatomical information relevant for understanding normal brain function (e.g. by studying connections between areas thought to implement certain functions and/or cortical thickness in some areas thought to host some functions); to compare brain anatomy in non-clinical groups of participants (e.g. normal and poor adult readers); and to study network properties thought to support some functions.

(3–4) Clinical fMRI (3) and Clinical sMRI (4) studies with patient groups including studies of ageing and studies focused on developmental disorders in children. Many clinical MRI studies compared brain function or structure across controls and patients or measured the effect of aging by studying multiple age groups. Single group studies also tested groups of participants in various experimental conditions. In a few papers participants were healthy 'control participants' but the primary objective of papers was clinical research (e.g. testing the effectiveness of pain suppression). Such papers were categorized as 'clinical' papers. The most frequently studied diseases and conditions and associated median and mean participant numbers in clinically oriented papers are shown in Supplementary Table 3.

(5–6) Normative developmental fMRI (5) and sMRI (6) studies with typically developing children who were under the age of 18 years. Many of these studies compared brain function or structure across multiple age groups.

There were very few cognitive neuroscience sMRI (16 papers with 18 studies) and developmental sMRI (19 papers with 19 studies) and fMRI (25 papers with 25 studies) studies as compared with studies in the other 3 categories. Hence, data from these 60 papers (62 studies) were not considered for analysis. However, the extracted data is available in the data file.

Data for the remaining 1038 papers with 1161 studies were analyzed in the work reported here. These studies were categorized as experimental fMRI, clinical sMRI and clinical fMRI studies. Table 1 shows the number of highly cited papers, the studies included in the papers and paper citation counts. Papers in this sample were published between

Table 1
The numbers of highly cited papers, the studies included in the papers and paper citation counts. The 1038 papers were subdivided into three categories (see details in text).

			Citation Counts				
Study Type	Papers	Studies	Min	Median	Mean	Max	Total
All	1038	1161	208	297	377	4147	391,178
Experimental fMRI	591	692	208	305	391	4147	231,071
Clinical sMRI	318	334	208	287	358	2912	113,954
Clinical fMRI	129	135	209	300	358	981	46,153

1990 and 2012 (Experimental fMRI studies: 1993–2012; Clinical sMRI studies: 1990–2011; Clinical fMRI studies: 1996–2012). The 1038 papers received 391,180 citations. The experimental fMRI papers received nearly 60% of these citations.

We extracted the following data for each study: 1) Total number of participants tested. 2) The number of participants stated as excluded from analyses. When no exclusions were reported we assumed that the number of excluded participants was zero. 3) The final number of participants included in the MRI analysis. This 'final' participant number was considered the number of participants for a study. 4) We determined whether a study defined two or more groups of participants. If at least two groups were defined then we recorded the number of participants in each group. 5) For experimental studies we noted the approximate main topic of a paper. 6) For clinical studies we coded the type of disease examined. 7) Finally, we coded whether a study was a randomized control trial or not.

2.2. Analysis of trial numbers in highly cited experimental fMRI papers

In order to get an impression of total and per condition experimental trial numbers in individual participants we have examined the Methods sections of 142 experimental fMRI studies with event-related designs where trial numbers should be well-defined in principle. We extracted the total number of trials and the number of experimental conditions where this was possible.

2.3. Sample of experimental fMRI papers in 2017 and 2018

In order to be able to compare sample sizes in the most highly cited papers to the most recent sample sizes at the time of data collection we have extracted data from a sample of papers from the two most recent years before writing this paper. Specifically, we have analyzed a sample of 131 experimental fMRI papers published during 2017 and 142 papers published during 2018 in 4 prominent neuro-imaging journals: *Nature Neuroscience, The Journal of Neuroscience, NeuroImage and Cerebral Cortex*. These journals were selected because they publish a large volume of experimental fMRI papers with normal adult participants *and* they were also in the top 7 most frequently occurring journals in our highly cited paper sample (see Supplementary Table 1). Hence, we could assume that many papers from these journals will become relatively highly cited in the future. Therefore, this sample is fairly complementary to the to the sample of highly cited papers. Obviously, for recent papers, as those published in 2017–2018, it will take several more years to determine which ones exactly will be highly cited.

We included about an equal number of papers from roughly similar issues in both 2017 and 2018, simply based on the availability of issues till the late summer of 2017 and 2018 (data collection periods). In order to determine eligibility for inclusion, the titles, abstracts and where necessary, the full text of the papers were briefly examined manually. The only inclusion criterion was whether a paper reported an empirical adult fMRI experimental study as defined in the highly cited paper sample. There was no other evaluation of content or other data from the papers before the final analysis. The number of studies and papers are shown in Table 2. The issues checked per journal are shown in Supplementary Table 4.

Table 2
The number of papers and studies in the 2017 and 2018 sample.

	2017		2018		Totals	
Journal	Papers	Studies	Papers	Studies	Papers	Studies
Nature Neuroscience	4	5	2	2	6	7
The Journal of Neuroscience	42	47	33	38	75	85
NeuroImage	46	51	66	73	112	124
Cerebral Cortex	38	44	39	40	77	84
Totals	130	147	140	153	270	300

During data collection we manually opened each pdf file in Adobe Acrobat © and checked the 'Participants' or equivalent section for initial and final sample sizes and for the number of excluded participants. In addition, we have also searched papers for the words 'power' and 'sample size' and determined whether papers included any formal power calculations and/or they justified their sample sizes. In order to verify power analyses in papers and/or to confirm their nature we have carried out our own power analyses for each paper based on the data given in the papers. This procedure was most often necessary because from papers it was often unclear what kind of power analysis was exactly done.

2.4. Data availability

All data and the analysis code (Matlab scripts; www.mathworks.com) producing all figures, tables and numerical details reported here are available at https://osf.io/qzerc/. The code produces figures that can be zoomed in and out in Matlab. It is not possible to upload pdf copies of published papers because of copyright restrictions (we have accessed papers through the subscriptions of the University of Cambridge, UK).

2.5. Ethics statement

This work did not test human or animal participants. All data was collected from published papers. Hence, no ethical permission was needed.

3. Results

3.1. Highly cited paper sample: 1990–2012

Fig. 1A compares the cumulative sample size distributions for the 3 types of papers in the highly cited sample. Fig. 1C sows corresponding histograms. Table 3 shows the number of participants in studies with a single group and with more than one group. For example, 662 out of 692 experimental studies had a single group of participants whereas 30 studies defined two or more groups. The median number of participants in studies with a single group was 12. In the 30 studies with groups the average group number was 2.033. The median number of participants in groups was 11. The proportion of studies with a single group is notably higher in experimental fMRI studies (662/692 = 0.9566) than in Clinical sMRI (163/334 = 0.4880) and Clinical fMRI studies (28/135 = 0.2074).

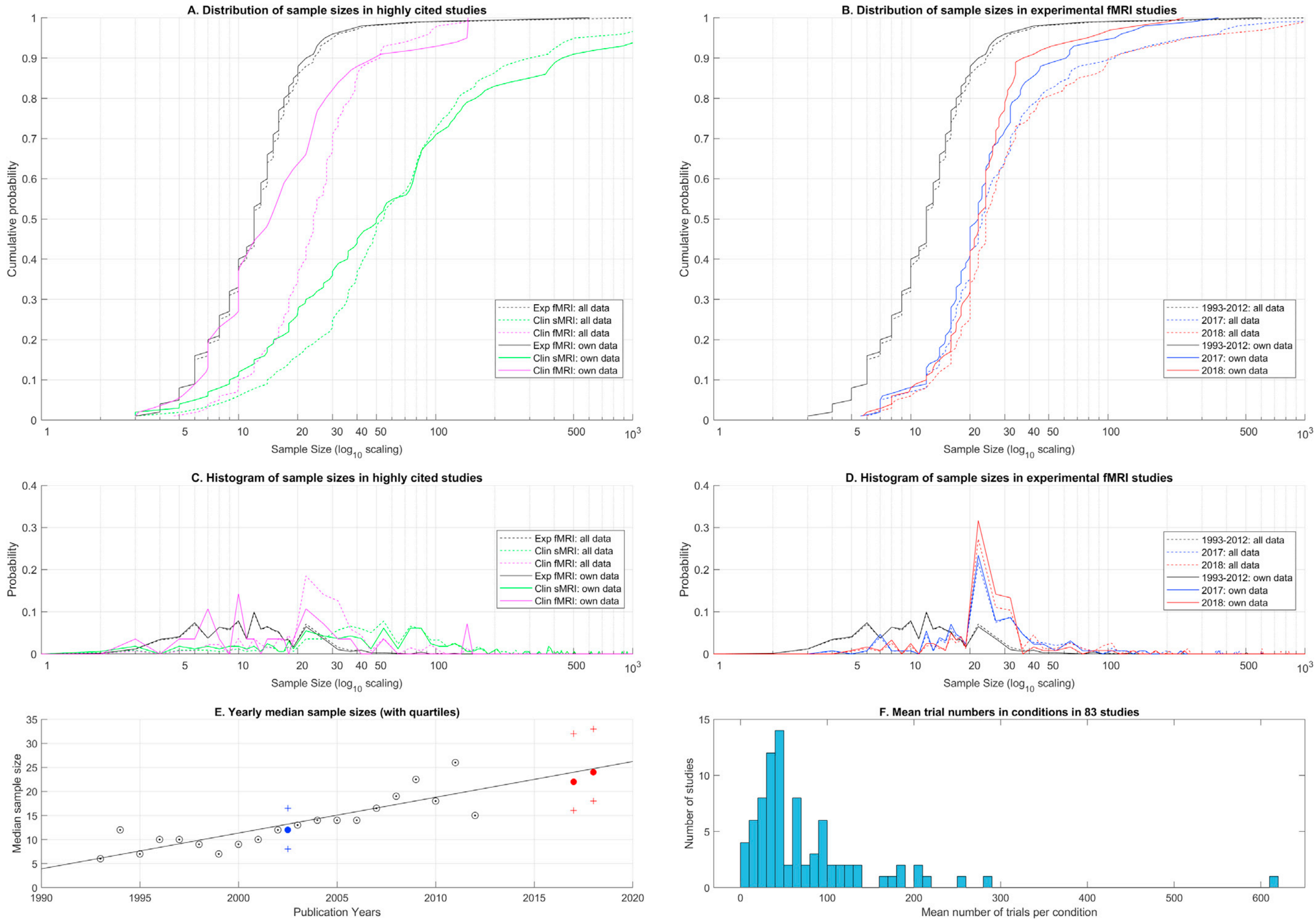


Fig. 1. **(A)** Cumulative sample size distributions in highly cited papers. The sample size/probability values are plotted for each sample size percentile. Dashed lines (-all data) in the legend show data for all studies. Continuous lines (-own data) in the legend show data restricted to studies that collected their own data and only included a single group of participants. The sample size axis is truncated at 1000 for better visibility. **(B)** Cumulative sample size distributions in experimental fMRI papers only (highly cited papers and 2017 and 2018 data). The sample size/probability values are plotted for each sample size percentile. Dashed (-all data) and continuous lines (-own data) show data as noted for Panel A. **(C)** Histogram of sample size distributions in highly cited papers (histogram bins: Each integer between 1 and 19; steps of 5 between 20 and 45 and steps of 10 from 50). **(D)** Histogram of sample size distributions in experimental fMRI papers only (histogram bins are the same as in Panel C). **(E)** Yearly median sample sizes. Black circled dots show the yearly medians of the sample sizes from the highly cited papers. The black line is the regression line fitted to this data. The leftmost blue dot and blue crosses represent the median and 25th and 75th percentiles of sample sizes from the entirety of the highly cited paper data. The rightmost two red dots and crosses represent the medians and 25th and 75th percentiles of 2017 and 2018 data. **(F)** The distribution of mean number of trials in the experimental conditions of 83 highly cited experimental fMRI papers (see further explanation in text). **(X)** Study categories in panels A–D: Experimental (Exp) fMRI and Clinical (Clin) sMRI and fMRI studies, 2017 and 2018 experimental fMRI studies.

Table 3
The number of participants in studies with a single group (gr = 1) and with more than one group (gr > 1). Study categories: Experimental (Exp) fMRI and Clinical (Clin) sMRI and fMRI studies. The first 3 columns show the number of studies with one or more groups and totals. The next 3 columns (N in Group if gr = 1) show participant numbers for studies with a single group. The next 3 columns (Number of Groups if gr > 1) show the number of groups in studies with more than one group. The last 3 columns (N in Groups if gr > 1) show participant numbers in groups in studies with more than 1 group.

	Studies			N in Group if gr = 1			Number of Groups if gr > 1			N in Groups if gr > 1		
Study Type	gr = 1	gr > 1	Total	Min	Median	Max	Min	Mean	Max	Min	Median	Max
All	853	308	1161	2	13	3660	2	2.35	8	1	17.2	1056
Exp fMRI	662	30	692	2	12	603	2	2.03	3	1	11.0	500
Clin sMRI	163	171	334	2	50	3660	2	2.50	8	2	24.0	1056
Clin fMRI	28	107	135	3	14.5	146	2	2.19	7	1	12.5	68

Median participant numbers were 3.5–4.17 times larger (N = 50) in single group Clinical sMRI studies than in the other two study categories. Median participant numbers were about twice as large in multi-group Clinical sMRI studies (N = 24) than in other study categories.

There was no statistically significant correlation between the number of citations to a paper and the number of participants in studies. In the whole sample of 1161 studies the correlation of citation count and sample size was r = 0.0098 [95% CI: -0.0413; 0.0609; p = 0.71]. In the sample of Experimental fMRI papers the correlation was r = 0.0342 [95% CI: -0.0405; 0.1084; p = 0.37].

3.2. Studies from 2017 to 2018

Table 4 shows the counts and proportion of papers and studies with their own data, with secondary data and with only one group in the 2017 and 2018 data sets. Similarly to the highly cited paper data the overwhelming majority of papers had a single group of participants. In 2017, of the 9 studies with secondary data 8 studies were single group studies, all of them had sample size ≥100. In 2018, of the 19 studies with secondary data 17 were single group studies, 10 had sample size ≥100. Studies with secondary data used various data sources, the most frequent source being the Human Connectome Project (humanconnectomeproject.org). In 2017 six out of nine while in 2018 seven out of 19 papers took their data from various editions of this data source.

3.3. Comparison of sample size percentiles from highly cited papers and from papers published in 2017 and 2018

Fig. 1B compares sample size distributions in the highly cited paper sample and in the 2017–2018 sample. Fig. 1D shows corresponding histograms. The shift in sample size distributions and a slight increase in the proportion of studies with large sample sizes (many of them studies based on data from large third party data bases) is well visible (sample size quantiles are shown in Supplementary Table 5). Table 5 shows what number and proportion of studies with their own data and a single group of participants exceeded certain 'landmark' participant numbers identified in previous studies (see Discussion for details).

Fig. 1E shows the increase in median sample sizes of experimental fMRI studies from 1993 to 2018. The figure shows the regression line fitted the medians (black dots) of highly cited experimental fMRI studies. The rate of increase in these medians was +0.74 participants/year (intercept = −1477). The blue dot (between 2002-3) shows the overall median and the small blue crosses show the 25th and 75th percentiles of the entire set of highly cited experimental fMRI study data. The two red dots and crosses on the right show the medians and percentiles for the 2017 and 2018 data. It is notable that extrapolation of the regression line extremely well fits the medians measured in 2017 and 2018. In Fig. 1E it is visible that medians show larger scattering relative to the regression line at the left and right extremes of data points. This is due to the fact that less data points were available in very early and very recent publication years (see numbers in Supplementary Table 6.) (Note that incidentally, this also illustrates larger variability in case of small sample sizes.)

3.4. The number of reported excluded participants in highly cited experimental fMRI papers

Table 6 shows the number of *reported* excluded participants per study category not considering studies with secondary databases (see below). 86–90% of highly cited studies did not report excluded participants. In contrast, 49% of studies in 2017 and 2018 reported some excluded participants. The proportion of studies with a relatively large number of excluded participants (6–10 or more excluded participants) notably increased from 1 to 2% in highly cited studies to about 6–11% by 2017 and 2018. More excluded participants were reported in clinical than in experimental studies.

In 2017 nine studies used data from secondary databases and none reported exclusions. In 2018 nineteen studies used data from secondary databases and 3 studies reported exclusions: 19, 206 and 306 participants. These studies were not included in Table 6 as their inclusion would largely inflate the mean number of excluded participants calculated there (to mean = 15.1).

Table 5
The number (N of studies) and proportion (Prop.) of studies with their own data and with a single group of participants that had sample size larger than or equal to 12, 24, 40, 80 and 100 (N ≥). Study categories: Experimental (Exp) fMRI and Clinical (Clin) sMRI and fMRI studies, 2017 and 2018 experimental fMRI studies.

Study type		N ≥ 12	N ≥ 24	N ≥ 40	N ≥ 80	N ≥ 100
Exp fMRI	**N of studies**	372	62	18	7	4
	Prop. (all = 662)	0.562	0.094	0.03	0.0106	0.006
Clin sMRI	**N of studies**	141	113	92	62	48
	Prop. (all = 163)	0.865	0.693	0.56	0.3804	0.294
Clin fMRI	**N of studies**	16	8	3	2	2
	Prop. (all = 28)	0.571	0.286	0.11	0.0714	0.071
2017	**N of studies**	117	53	20	8	7
	Prop. (all = 127)	0.914	0.414	0.16	0.0625	0.055
2018	**N of studies**	109	57	12	6	4
	Prop. (all = 120)	0.908	0.475	0.1	0.05	0.033

3.5. Trial numbers in a sample of highly cited experimental papers

We could identify total and per condition trial numbers in 109 of the 142 experimental fMRI papers with event-related designs, while this information was unclear in the other 33 papers.

17 papers described old/new recognition memory experiments. We extracted the number of memory encoding trials as the usual question of interest is whether some brain activity at encoding will predict later recognition. The number of trials ranged from 10 to 455 (median = 150). 9 papers described designs with a large number of standard trials interspersed with a significantly lesser number of deviant trials from a different, critical trial type, for example in go/nogo designs (where typically there are many fewer nogo than go trials) and in task switch designs (where typically there are much fewer task switch than standard trials). Trial numbers varied from 128 (8 critical) trials to 1180 (80 critical) trials.

In 83 of the 109 papers trial numbers were more similar across conditions than in the above standard/deviant like designs. However, trial numbers were still very often unequal across conditions and there was also great variability in design. Fig. 1F shows the mean number of trials by condition. The total number of trials in an experiment ranged between 40 and 2440, the number of conditions ranged between 2 and 28 and the

Table 4
The counts and percentages of papers and studies with their own data, with secondary data and with only one group in the 2017 and 2018 data sets. The studies with one group are a subset of the studies with their own data. Percentages are computed relative to total paper and study numbers as shown in Table 2 (e.g. 121 + 9 = 130 papers in 2017 and 122 + 18 = 140 papers in 2018).

	2017				2018			
Data source in study	Papers	%	Studies	%	Papers	%	Studies	%
Own Data	121	93.1	138	93.9	122	87.1	134	87.6
Secondary data	9	6.9	9	6.1	18	12.9	19	12.4
Own data and Single group	110	84.6	127	86.4	108	77.1	120	78.4

Table 6

The number of excluded participants per study category in studies with their own data. Numbers are given for studies (not papers). Study categories (table rows) for highly cited data: Experimental fMRI and Clinical sMRI and fMRI studies. The 2017, 2018 data included experimental fMRI studies. For each category the following data are shown (table columns): total number of studies (Total), studies with no reported exclusions (None), studies with some exclusions (Some), studies with certain numbers of participants excluded (1–5, 6–10, >10). The median and mean of the number of excluded participants. For each category, top rows (N) communicate study numbers and bottom rows (%) communicate the percent of studies relative to the total study numbers shown here (The 9 studies in 2017 and the 19 studies in 2018 studies with secondary databases are omitted from this table; see text for details).

		Total	None	Some	1–5	6–10	>10	Median	Mean
Experimental fMRI	N	692	595	97	80	13	4	2	3.7
	%	100	86	14	12	2	1		
Clinical sMRI	N	334	301	33	6	7	20	22	87.5
	%	100	90	10	2	2	6		
Clinical fMRI	N	135	121	14	8	3	3	4	8.4
	%	100	90	10	6	2	2		
2017	N	138	70	68	47	13	8	3	4.985
	%	100	51	49	34	9	6		
2018	N	134	68	66	41	15	10	3	6.2
	%	100	51	49	31	11	7		

Table 7

Summary of power calculation results. The table shows data for the 130 papers in 2017 and the 140 papers in 2018. It is shown whether papers included statistical power calculations (a), had any comments on power (b) or had no comments on power (c). Subcategories of papers with power calculations are also shown a priori and post-hoc power computations and cases where this could not be determined. The numbers adding up to the total numbers of papers in a year are in italics: 9 + 33+88 = 130; 9 + 43+88 = 140.

	2017		2018	
Information on power?	N	% of 130	N	% of 140
Power calculation (a)	*9*	6.9	*9*	6.4
- A priori power calculation	4	3.1	6	4.3
- Post-hoc power calculation	3	2.3	3	2.1
- Unclear	2	1.5	–	–
Comments on power (b)	*33*	25.4	*43*	30.7
No comments on power (c)	*88*	67.7	*88*	62.9

mean number of trials per condition ranged between 4 and 610. For example, on the one extreme 112 trials were distributed into 28 conditions and on the other end 2440 trials were distributed into 4 conditions. It is notable in the figure that the mean number of trials per condition tends to decrease as the number of experimental conditions increases.

3.6. Overview of power calculations in the 2017 and 2018 papers

Table 7 shows the summary of the statistical power analysis assessment. In both 2017 and 2018 less than 7% of papers (9 papers in both years) included power calculations and about a third of the papers made some comment about power. None of the papers with large secondary databases had power calculations in any of the years. 7.6% (10) vs. 11.3% (16) of papers without power calculations referred specifically to the problem of having low power, in 2017 and 2018, respectively. Only 3–4% of 2017 and 2018 papers had clearly a priori power calculations.

3.7. Details of power calculations in the 2017 and 2018 papers

In 2017 four papers seemed to include formal pre-study power calculations ($\alpha = 0.05$ for all). Two of these papers ($n = 36$ and 53) computed power for single runs of t-tests (two cases; Cohen'D = 0.5 and 0.65; power = 0.8 for both). Two other distinct papers ($n = 32$) in the same journal issue used the same participants. One paper computed power for a single one-sample *t*-test ($D = 0.5$) and set a priori power to exceed 0.85. The other computed power for a single product-moment correlation ($r = 0.4$) and set a priori power to exceed 0.75.

In 2018 six papers described a-priori power computations. *One paper* determined that a sample size of 24 was necessary to achieve power = 0.8 with a matched-sample two-tailed *t*-test to detect an effect size of $D \geq 0.6$ ($\alpha = 0.05$). However, 2 participants were excluded from analyses leaving only 22 participants in the study. This obviously left the study underpowered by its own power criterion. *Another paper* aimed to look at brain structure vs. behavioral performance correlations ($n = 40$). It was not specified how exactly power computation was done but our own power analysis suggested that the required sample size of 34 was computed for power = 0.8 for a single correlation test of $r = 0.4$ ($\alpha = 0.05$; one-tailed). The study noted that power was computed for $D = 0.4$ in G-Power (Faul et al., 2007). However, G-Power computes the sample size of 34 when $r = 0.4$ rather than when $D = 0.4$ (entering r is the default in G-Power). For $D = 0.4$; r would be $r = D/\mathrm{sqrt}(D^2 + 4) = 0.1961$ (Borenstein et al., 2009). To detect an effect size of $r = 0.1961$ G-Power computes that $N = 156$ participants would be necessary ($\alpha = 0.05$; one-tailed). For this effect size ($r = 0.1961$) the study would have achieved power = 0.31 with $N = 34$ and actually achieved power = 0.34 with 40 participants ($\alpha = 0.05$; one-tailed). Hence, the study had much less power than reported. Alternatively, it is possible that the power computation parameters were misreported, and the intention was to compute power for $r = 0.4$. *Another paper* stated that they chose a sample size to achieve good power to detect the typical effect size in the field. An effect size of $r = 0.54$ was chosen from a previous meta-analysis. It was concluded that for the 25 participants initially tested power = 0.87 would be achieved for the $D = 0.54$ effect size ($\alpha = 0.05$; two-tailed; power was computed for a single correlation). The paper initially tested 25 participants but one participant was excluded, so only 24 participants were tested.

Another 2018 paper computed that 34 participants were necessary to detect an interaction effect size of partial eta^2 = 0.06 at power = 0.8 and 12 participants were required to replicate a previously found effect size of partial eta^2 = 0.17 at power = 0.8. The study tested 34 participants. Based on recomputing power in G-Power it seems that power was computed for a 2(groups) x 2(measurements) mixed design ANOVA with partial eta^2 of 0.06 and 0.17 (transformed to Cohen's $f = 0.2526$ and $f = 0.4225$ by G-Power). *Another paper* referred to a previous study of the authors where they used Monte-Carlo simulation to estimate sample size. The current study aimed to test a sample size of 30 participants as required by this previous power calculation. Based on the calculation the authors concluded that they achieved power >0.95. However, 2 participants were excluded, so only 28 participants were tested. *One paper* estimated the required sample size for Multi Voxel Pattern Analysis by a simulation method ($n = 87$).

In 2017 in two cases it was unclear whether power was computed a priori. In these papers power (set to 0.8) was computed for a *t*-test ($D = 0.44$; $n = 128$) and for correlation ($r = 0.5$; $n = 29$). The first of these papers presented the only RCT in our 2017 sample.

In 2017 in three cases power was computed post-hoc. In these papers power was computed for a *t*-test ($D = 0.6$; $n = 22$), for an ANOVA interaction term ($D = 0.52$; $n = 2 \times 15$). The latest study computed power to guide future studies. In one case ($n = 8$) it was unclear how power was computed as no exact effect sizes were given (but likely for multiple t-tests).

In 2018 power computations seemed clearly post-hoc in three cases. *One paper* with 15 participants has achieved null results in whole brain analyses. The study has computed the achieved power for multiple testing uncorrected ROI analyses suggesting that analyses achieved power = 0.8 to detect an effect size of D = 0.68 with α = 0.05 and D = 0.56 at α = 0.1. Computations were not specified in the paper but reanalysis suggests that they were done for one-tailed t-tests. *The second paper* noted that the study was 'adequately powered to detect large effects' and noted that the sample size of 12 was adequate to detect an effect size of d ≥ 0.89 at power = 0.8 (α = 0.05). Power was computed for a single two-tailed matched-sample *t*-test. *In the third paper* it was unclear how power was computed but the authors claimed to have run some analyses for effect sizes of D = 0.51, 0.53 and 0.54 with power = 0.8. The analyses close to describing the power computation mentioned the use of matched sample t-tests and the study had 31 participants. Indeed, power for a two-tailed *t*-test (α = 0.05) for the above effect sizes and sample size varies between power = 0.78 to 0.83. So, it is likely that power was computed for single matched-sample t-tests.

Besides the papers with power calculations 34 and 44 papers mentioned statistical power in 2017 and 2018, respectively. Mentions were most often non-specific and non-informative. For example, only one paper in Nature Neuroscience had power calculations but all of them included similar text stating that no methods have been used to predetermine sample sizes and sample sizes were simply chosen to be in line with common practices in their field. Many studies noted that their sample size was chosen so that they would be identical to or exceed sample sizes from the authors' own or others' previous work. Some noted that sample size was based on funding availability. Some studies commented on issues of statistical power in general without clearly linking it to the context of the specific study. Some studies commented that a certain analysis was less or more powered than another one without giving any further details or computations. Some papers commented on their large perceived sample size (e.g. 60) without giving any actual power calculation details. Ten (2017) and sixteen (2018) studies mentioned that they may have been underpowered. Usually these non-informative statements were restricted to one or two brief comments in a paper. Several studies used small subsamples from their overall sample for certain analyses.

In 2017 seven whereas in 2018 only two papers had multiple studies where some of these studies declared a goal to replicate findings from an earlier study in the same paper. None of these papers had power calculations.

In 2017 two studies were special cases focused on individual measurement: One had only 4 participants but each of them were tested in 5–6 sessions. The other tested 10 participants, each for 300 min during 10 sessions.

3.8. The prevalence of NHST statistics in papers

Power analyses are important for NHST statistics. Hence, we can ask the question in what proportion of the above papers NHST statistics were used at all. In order to gain an impression of this we have run an automated text search for all papers in the 2017 and 2018 sample. We searched for the following terms strongly associated with NHST statistics in neuroimaging: 'ANOVA', 'ANCOVA', '*t*-test', '*t* test' [with space separator], 'p = ', 'p<', 'p>' and 'p≤'. Note that this list is not exhaustive, for example, we have not searched for the expressions 'significant', 'significance', 'correlation', etc. as their interpretation could be ambiguous. So, our text search was fairly conservative. In addition, the p value search terms could not pick up all p values due to journal specific character coding issues. So, there were likely many more p values reported than picked up by the algorithm.

The text lines extracted around search terms (the search term + 50 characters) are available as Supplementary Data File 1. An examination of this data suggests that the appearance of search terms was virtually always associated with the corresponding NHST statistics used in papers. In summary, we found that at least one of the search terms appeared in about 95% of papers in both 2017 (123 of 130 papers) and 2018 (134 of 140 papers). We have manually searched the 7 papers in 2017 and 6 papers in 2018 in which the search terms were not found. Only one single paper did not report p values and NHST statistics. Hence, we can conclude that >99% of papers in our sample used NHST statistics.

4. Discussion

Running underpowered studies may waste research funding on studies which a priori have low chance to achieve their objectives. In addition, low power leads to high false report probability, imprecise measurements (and therefore highly variable findings in terms of effect sizes and spatial/anatomical location) and effect size exaggeration. Here we have shown that participant numbers are relatively low in the most highly cited fMRI papers. Such low sample sizes have been associated with low statistical power for typical effect sizes in this field (Desmond and Glover, 2002; Murphy and Garavan, 2004; Yarkoni, 2009; Ingre, 2013; Lindquist et al., 2013; Ioannidis, 2008, 2005a,b; Button et al., 2013; Poldrack et al., 2017; Szucs and Ioannidis, 2017a,b; Turner et al., 2018; Geuter et al., 2018). Hence, highly cited studies are likely to have similar problems stemming from low statistical power as most 'typical' neuroscience studies. Our analysis also shows that sample sizes are slowly but steadily increasing. However, the rate of increase is low. We also found that power calculations are exceedingly rare in the published literature.

4.1. The number of participants per group

Highly cited experimental and clinical fMRI studies had similar median sample sizes (medians in single group studies: 12 and 14.5; median group sizes in multiple group studies: 11 and 12.5). 96% of experimental studies were single group studies. This pattern remained in 2017 and 2018 when 93% and 87% of experimental fMRI studies had a single group. Single group studies most likely had within-subject designs testing the same participants in two or more conditions. While within-subject designs are more powerful than between subject designs, even a single one-sample *t*-test (two-tailed) requires 34 participants to surpass 80% power to detect an effect size of D = 0.5 at α = 0.05. Of note, many power calculations we found (optimistically) expected a similar effect size that would be quite substantial in behavioral research. More than 90% of highly cited studies had a smaller sample size than 34 and even in 2017/2018 more than 80% of studies remained under this sample size. If the α level is decreased to α = 0.001 to be more in line with fMRI standards, then detecting a D = 0.5 effect size with 80% power would require 74 participants and detecting a D = 0.3 effect would require 196 participants (90 participants at α = 0.05). At the α = 0.001 level even detecting an effect size of D = 0.8 would require 33 participants (with 80% power). While the above power calculations are illustrative, fMRI specific power calculations have also convincingly demonstrated that typical participant numbers in our sample have very low power (Turner et al., 2018; Geuter et al., 2018). For example, Geuter et al. (2018) concluded that to detect an effect size of D > 0.8 requires more than 40 participants while to detect an effect size of 0.5 < D < 0.8 requires a sample size of at least 80. Turner et al. (2018) also found that a sample size of 36 assures very low replicability and optimizing replicability requires sample sizes well beyond 100. Clearly, most studies in our data had much lower sample sizes than these values.

In contrast to experimental studies, only 21% of highly cited clinical fMRI studies were relatively small single group studies. However, while clinical fMRI studies had somewhat larger individual participant groups than experimental studies (11 vs. 12.5, see above), most of their total sample size advantage stemmed from the fact that they more often had two or more groups than experimental studies. This is important to consider when comparing sample sizes from clinically vs. non-clinically oriented journals/publications. Clinical studies included multiple groups because these studies often included both patient and control groups. Overall, group sizes were very similar in both experimental and clinical fMRI studies. Importantly, for the same overall sample size independent sample t-tests are less powerful than one-sample or matched sample-tests often used in single group studies (see e.g. GPower; Faul et al., 2007). Hence, in terms of group comparison clinical fMRI studies probably do not have a power advantage over experimental fMRI studies. Only the single-group clinical sMRI studies had notably larger sample sizes than fMRI studies. The discrepancy between fMRI and sMRI studies probably has to do with the extra time and effort necessary to collect and analyze fMRI than sMRI data. Genuine effect sizes may be larger in clinical than in experimental studies because disease is likely to have more substantial impact on brain function than experimental manipulations in healthy participants. Hence, clinical studies may have some power advantage due to larger expected effect sizes than in experimental studies.

4.2. *Growth in sample sizes*

We found that median sample sizes in highly cited experimental fMRI studies increased consistently at a rate of +0.74 participant/year between 1993 and 2010. This rate of increase was perfectly in line with the median sample sizes we found in 2017 (23) and 2018 (24). The +0.74 participant/year rate of increase was also in line with our survey (Szucs and Ioannidis, 2017a) examining 3801 papers published between 2011 and 2014. In this earlier paper we reported degrees of freedom for one or two-sample t-tests and estimated that the median degree of freedom was 18 in cognitive neuroscience papers. Provided that median sample sizes were likely to be about 1–2 larger than the degrees of freedom this data would also well fit the regression line found in the current study (with about median sample size of 19–20 in about 2012/13). Notably, our sample size median estimates are smaller than the 28.5 median estimated by Poldrack et al. (2017) for the year of 2015. However, our analysis is well compatible with the full set of data points from David et al. (2013) used by Poldrack et al. (2017).

Overall, our current and earlier data (Szucs and Ioannidis, 2017a) and data from other evaluations (David et al., 2013; Poldrack et al., 2017) suggest that sample sizes and consequently, power are improving, albeit very slowly. Only ~10% of highly cited experimental fMRI papers published between 1993 and 2012 reached the sample size of 24 suggested by Desmond and Glover (2002) and only about 3% reached the sample size of 40 that Geuter et al. (2018) considered adequate to detect only large effects. There has been clear improvement by 2017 and 2018 when respectively, 41% and 48% of papers reporting their own data and with a single group of participants (constituting about 90% of all papers) were above the minimum participant numbers recommended by Desmond and Glover (2002) eighteen years ago. However, this also means that in 2017 and 2018 still more than half of these papers had less than 24 participants. Moreover, there has been less improvement at the higher end of participant numbers as in 2017 and 2018 still only 16%, 6% and 5% of the above papers had more than 40, 80 and 100 participants, respectively (Geuter et al., 2018; Turner et al., 2018). These proportions were 10%, 5% and 3% in 2018. Notably, besides the α level, power depends on the sample size and the effect size searched. Hence, studies cannot universally rely on absolute sample size guidelines. An optimal approach may be to determine a sample size that is appropriate to detect a 'theoretically informative' minimum effect size (Poldrack et al., 2017).

Importantly, both here and earlier (Szucs and Ioannidis, 2017a) we detected considerable variability in sample sizes across studies. So, using solely medians to characterize sample sizes seems inadequate as it masks substantial variability. The crucial question is what proportion of studies in the literature remain too small and hence, underpowered.

While we only have two years' worth of observations from 2017 to 2018 a noteworthy trend in the literature is the increasing use of large third-party databases in neuroimaging. The proportion of papers using such databases doubled from 2017 to 2018 from 6% of studies to 13% of studies. The use of these databases more than doubled the number of studies with more than 100 participants. It remains to be seen whether this trend continues in the coming years and whether it is present in other neuroimaging journals. The use of large shared databases would be beneficial for many reasons. First, such databases assure high statistical power for modest effects. Second, considering the effort required to compile large databases data collection may be carried out by seriously vetted procedures and by experienced teams. Third, data is available to any interested researchers assuring increased scrutiny and hence, reliability of published results. Fourth, if data is collected in a decentralized manner (e.g. many labs jointly contributing to data collection) than replicability across different labs can easily be examined. Of course, there are also downsides of using only large databases: e.g. biases may recur in the literature if a large and often used database is inherently biased in some ways and the number of false positive errors may accumulate due to sequential multiple testing by many researchers (Thompson et al., 2019). For example, here we found that data from various editions of the Human Connectome Project were used in close to 50% of studies (13 of 28 studies) using secondary datasets in 2017–2018. Further, it is also possible to search large datasets for a subset of data confirming some predefined hypotheses, or apply machine learning procedures searching for patterns in noise (Powell et al., 2020). Hence, it is important to use data from secondary databases in a principled manner, for example, using all available and appropriate data (rather than just a subset) for hypothesis testing clearly defining exclusion principles.

4.3. *Power calculations*

While low power in neuro-imaging received lots of attention recently (Poldrack et al., 2017; Szucs and Ioannidis, 2017; Button et al., 2013), we found that in both 2017 and 2018 only about 3–4% of papers had clear pre-study power calculations and more than 62% of papers never mentioned any issues of statistical power. Most power calculations we found were done for single runs of t-tests and product-moment correlations as there seems to be no agreement on how to estimate statistical power for fMRI studies which rely on a very large number of tests, idiosyncratic statistical procedures and on heavy multiple testing correction (Hayasaka et al., 2007; Poldrack et al., 2017; Carp, 2012). The power calculations we found often expected medium sized effects based on previous published data. However, considering the very probable effect size inflation of the published literature expecting relatively large medium sized effects seems too optimistic (Ioannidis, 2008; Szucs and Ioannidis, 2017a). It also frequently happened that studies determined a required sample size by power calculation but then analyzed less data than required by their own power calculation because they did not account for the number of excluded participants. This practice leaves studies underpowered by their own power criteria. It was also typical that power calculation parameters were not defined clearly so that in most cases guesswork and recalculation was necessary to see how power was determined. In some cases power calculations seemed erroneous.

Many papers without power calculations referred to sample sizes in previous similar research to justify their sample sizes. However, considering that lots of neuroimaging is underpowered (Yarkoni, 2009; Button et al., 2013; Szucs et al., 2017a) this is clearly inadequate rationale. Unless the purpose is to guide future studies it is not informative to compute post-hoc power considering an effect size already detected as statistically significant in a study. Moreover, small studies are not good guides for power calculations for future studies because they can *only* detect relatively large effects as statistically significant (see e.g. Ioannidis, 2008; Yarkoni, 2009; Szucs and Ioannidis, 2017a,b; Gauter et al., 2018). Similarly, meta-analyses may also overestimate effects because they tend to rely on many small, underpowered studies (Ioannidis, 2010). In fact, studies with large sample size (and hence, with more accurate measurements than small studies) rarely report large effects (see for example Fi. 2. in Szucs and Ioannidis, 2017a, and comments in Yarkoni, 2009).

A notable observation is the near complete lack of the use of fMRI specific power calculation procedures. While these procedures have been available since a while (e.g. www.neuropowertools.org; Durnez et al., 2016; see useful links to websites in Poldrack et al., 2017; Mumford and Nichols, 2008; Mumford, 2012; Gauter et al., 2018; Turner et al., 2018; Desmond and Glover, 2002; Murphy and Garavan, 2004) it seems that their use may be seen as being complicated, researchers may not know these procedures yet, or they may not consider the use of these procedures high enough a priority for investing effort. The power calculations that we found mostly focus on single *t*-test and correlation analyses and they likely overestimate statistical power as they do not consider multiple-testing correction at all. Finally, we observed that all but one paper in the 2017/2018 sample used some NHST statistics. Hence, the use of power analyses would have been justified in most papers. Our data is consistent with evidence from text mining of the entire biomedical literature that suggests that use of NHST statistics with p-values is the standard statistical approach (Chavalarias et al., 2016; Ioannidis, 2019).

4.4. High population level power vs. small N designs

A further point to discuss regards the question of whether high population level power is always necessary for studies. As the overwhelming majority of fMRI papers are making population level claims about the precise location and/or quantitative aspects of brain processes (mostly related to localizing some function in the "general human brain") we suggest that high population level power is necessary in most studies. This is even more important if studies claim clinical relevance. If studies are so vastly underpowered, claims cannot be generalized to a wider/selected population and can only be evaluated for the group of participants tested (Friston, 1999).

A related note concerns the use of so-called small N designs (Smith and Little, 2018). We agree that small N designs can be more appropriate than large samples in some situations. However, *simply* having small N does not turn a study into a *credible* small N study. Indeed, the use of small-N designs is optimal if both measurement and theory (precise quantitative *models*) are strong and there is excellent measurement precision within participants (Smith and Little, 2018). Hence, small N designs require *extremely high-powered* individual measurement (e.g. delivering thousands of trials to participants in psychophysics experiments) and in fact typically aim to *replicate* their findings multiple times in a single participant and across a group of a few participants (Smith and Little, 2018). In contrast, in most neuroimaging studies theory is relatively weak (there are no clear quantitative predictions), measurement is weak (e.g. the expected effect size cannot be defined) and within-subject power is very weak (Szucs and Ioannidis, 2017b). For example, as shown in the Results here, currently very few studies deliver high trial numbers. Overall, it is a well-known problem of neuroscience experiments that it is often difficult to identify even well-established group level effects in individual participants. This is probably due to both high individual variability and to low individual and group level power that results in highly variable findings both in their spatial localization and effect size (see Gauter et al., 2018). Due to the above reasons currently most imaging studies cannot claim to have credible small N designs.

Previously we have argued that both neuroimaging and psychology should rely much less on 'blind' (aiming to reject a weakly motivated 'nil' null hypothesis) statistical hypothesis testing and should rely more on parameter estimation and building quantitative models for the observed data values (Szucs and Ioannidis, 2017a). The use of credible small N designs (Smith and Little, 2018) is in perfect agreement with this suggestion and may even address the problem that fMRI researchers do not have unlimited scanning resources. First, credible small N designs would allow us to build and test strong quantitative models in a few participants. Then, if our models already work well in a few participants, they could predict exact quantitative effect sizes at least in some Regions of Interest. Subsequently, if population generalization is important, larger preregistered studies with high statistical power could be run for a population of interest.

At the practical level, considering that signal to noise ratio is proportional to the square root of trials used in experiments assuring high individual level power would require collecting much larger volumes of individual data than currently typical. Using many trials/epochs would then substantially prolong experimental time for individual participants. Further, the number of trials in single fMRI measurement sessions is constrained by various practical factors. First, the sluggish nature of the haemodynamic response requires relatively long trial durations (e.g. much longer than in electro-encephalography experiments). Second, fMRI scanning time is expensive (e.g. costs may be higher than £500/hour in the United Kingdom). Third, lying in the scanner relatively motionless is tiring for participants. Fourth, complex cognitive experiments may need long trial durations which restrict the number of trials doable in a single imaging session. The only solution may be to collect many trials in multiple runs. However, in such a case multi-level analysis is needed to factor in potential discrepancies across sessions. Moreover, research grants can rarely offer funds for long/repeated scanning sessions. Hence, overall several practical limitations often beyond the control of researchers restrict increasing individual measurement precision in studies.

Regarding experimental trial numbers our data shows that individual research groups may use very different trial numbers even in relatively similar experimental designs. Overall there is very great variability in trial numbers per experiment condition in the literature. Hence, individual measurement precision is likely to vary greatly across studies and research groups. A note concerns the potential approach of aiming to increase confidence in group level findings by replicating initial findings from an experiment in a second pre-registered experiment within the same paper. For example, this approach could be used in explorative studies that could not initially include power calculations because relevant effect sizes could not be determined but they still intend to make population level claims. Clearly, without replication experiments such claims from explorative findings must be treated with the utmost caution and should not be considered 'scientific truth' (see discussion in Szucs and Ioannidis, 2017a). While the above approach would be beneficial, this practice is currently very rare: 9 out of 273 papers had replication experiments but none of these papers had power calculations. In addition, it may also be difficult to ensure independence of replication experiments reported within a single paper.

4.5. Participant exclusions

Only 10–15% of highly cited studies reported any excluded participants. In contrast, in 2017 and 2018, 49% of studies reported at least some excluded participants. The proportion of studies with more than 5 excluded participants also increased by about 5-fold by 2017/2018. Importantly, studies practically never stated that *no* participants were excluded. Hence, the default value was that studies have not mentioned anything about exclusions. That is, when there were no reported exclusions it may mean that there were really no exclusions or that exclusions were simply not reported. Taking the above into account our observations about reporting exclusions raise several questions. On the one hand, the distribution of excluded participants in highly cited papers seems oddly biased towards 0 exclusions and it is also in conflict with the much larger proportion of studies with exclusions and the larger number of excluded participants in 2017/18. So, many highly cited studies may not have reported exclusions rather than not have exclusions. If so, there may have been a change in exclusion reporting habits during the past years. Alternatively, perhaps the most recent papers do exclude more participants then earlier papers. For example, more recent papers may apply more stringent exclusion criteria than in the past by excluding participants with noisy measurements. Were this this case, the higher noise level of highly cited studies would result in more false positive outcomes than in more recent studies. Overall, it is not possible to decide which of the above scenarios may be more probable. However, it is important to note that exclusions should be well documented in all cases as arbitrary exclusions do allow for high 'researcher degree of freedom' and so have implications for data dredging and N-hacking (Simmons et al., 2011; Carp, 2012; Szucs, 2016). For example, if participants not confirming to group averages are simply deemed 'noisy' and are excluded from samples then the data will get seriously distorted and will not be representative of the population. Therefore, it would be important to clarify and define the *exact* exclusion criteria and numbers in research fields. If data is published, data from excluded participants should also be published.

4.6. The clarity of reporting

Besides the uncertainty about interpreting the above exclusion data it is also noteworthy that extracting condition numbers and trial numbers in each condition in published papers was particularly difficult due to completely idiosyncratic descriptions and missing information. For example, it is striking that in nearly 25% (33/142) of the highly cited event-related design experimental fMRI studies we were unable to find out how many trials were used in experiments. These results are consistent with previous observations on poor reporting standards in neuroimaging (Carp, 2012; Guo et al., 2014). Unclear reporting of study parameters is a particular danger in neuroimaging where a large number of often complicated and opaque procedures are used and seemingly minor changes in some (undocumented) (pre-)analysis parameters can result in major distortions of data (Carp, 2012). In order to increase reporting standards, the Organization for Human Brain Mapping's 2015-16 Committee on Best Practices in Data Analysis and Sharing (COBIDAS) has formulated several reporting guidelines (Nichols et al., 2016; Nichols et al., 2017; for a summary see Box 4 in Poldrack et al., 2017). We suggest that journal editors need to enforce these guidelines. Based on these reporting guidelines it would also be desirable to quickly develop 'industry standards' for the *exact* reporting *formats* for technical aspects of neuroimaging studies including power requirements. For example, the creation of the Brain Imaging Data Structure (BIDS; Gorgolewski et al., 2016) is an attempt to standardize data sharing/reporting formats. A complementary approach could be to link *standard reporting cards* to *all* neuroimaging papers. For example, Nature Research has already started using standard 'Reporting Summaries for MRI studies'. However, we suggest a more formal, comprehensive and universally required approach providing detail on all standard aspects of imaging studies (see related discussion in Begley and Ioannidis, 2015). Using such reporting cards would also make researchers' job easier as they could check whether they have carefully planned and documented all aspects of their study. Standardized reporting would also make papers easily machine readable, with their data being possible to re-analyze and to combine.

5. Conclusions

Besides clear and standardized reporting, in our opinion two key ingredients of future population-level hypothesis testing studies are pre-registration (optimally, with pre-study acceptance by journals) and a principled increase in sample sizes (Hardwicke and Ioannidis, 2018; Munafo et al., 2017; Poldrack et al., 2017; Szucs and Ioannidis, 2017b; Ioannidis et al., 2014). Pre-registration guarantees that studies get published based on pre-study significance and therefore may largely decrease incentives for rephrasing exploratory results to give the impression of pre-hypothesized and expected results. Decreasing publication bias would also decrease effect size exaggeration (as negative findings would get published). Principled sample size increase should be based on pre-study power calculations specifying sought after realistic effect sizes in NHST studies.

The consistent historic increase in sample sizes suggests that we may be able to break the long 'tradition' of criticizing low power but not improving the situation (Sedlmeyer and Gigerenzer, 1989). However, the increase in sample sizes could be sped up by targeted and timely interventions by both publishers and funders. Funding contracts could specify power-calculation-based sample sizes, fund studies appropriately if population generalizability is of interest, require pre-registration for studies with pre-defined hypotheses, standardized reporting of methods/results, the necessity of publishing all analysis code and raw data (Poldrack et al., 2017). Such changes would provide funders with certainty that their money is not wasted.

It is tempting to assume that many of the highly cited papers analyzed here are probably replicated given that so many other scientists cite them. However, high citations are not synonymous with replication. It is well known from other fields that some papers get extremely heavily cited without any attempt to replicate them and that when replication eventually is attempted, it fails (Ioannidis, 2007; for neuroimaging see: Boekel et al., 2015). A survey of the most-highly cited papers across all medicine has shown that of the most-cited observational studies 5 out of 6 were subsequently refuted and even a quarter of randomized trials were contradicted (Ioannidis, 2005). Exact replication in particular is often avoided and this may allow building large literatures upon questionable findings (Ioannidis 2007, 2012). Therefore, we suggest that, whenever this has not been done already, the *exact* replication of some highly cited influential studies should be high priority as many of these ground-breaking studies were done in a previous era with deficient sample size standards. Consequently, some highly cited studies in neuroimaging may have high false report probability (Szucs and Ioannidis, 2017a; b).

Author contributions

DS designed the research, extracted data from highly cited papers, analyzed data, wrote program code and the first draft of the paper. JPAI contributed critical comments to design, and data interpretation and revised successive drafts with DS.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

The study received funding from the James S McDonnell foundation through the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition - Scholar Award (DS; No 220020370). METRICS is supported by a grant from the Laura and John Arnold Foundation. The work of John Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell. We thank Timothy Myers for an initial check on sample sizes in about half the highly cited papers. We thank Josefína Weinerova for extracting sample size data for 2018. We thank Rik Henson (University of Cambridge) for comments on an earlier version of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.117164.

References

Amrhein, V., Greenland, S., McShane, B., 2019. Retire statistical significance. Nature 567, 305–307.

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in science: improving the standard for basic and preclinical research. Circ. Res. 116, 116–126.

Benjamin, D.J., Berger, J.O., Johannesson, M., et al., 2018. Redefine statistical significance. Nat. Hum. Behav. 2, 6–10. https://doi.org/10.1038/s41562-017-0189-z.

Boekel, W., et al., 2015. A purely confirmatory replication study of structural brain–behavior correlations. Cortex 66, 115–133.

Borenstein, M., Hedges, L.V., Higgins, J.P., Rothstein, H.R., 2009. Introduction to Meta-Analysis. John Wiley and Sons, Ltd.

Button, K.S., Ioannidis, J., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafo, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.

Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. Neuroimage 63, 289–300.

Chavalarias, D., Wallach, J.D., Li, A.H., Ioannidis, J.P., 2016. Evolution of reporting P values in the biomedical literature, 1990-2015. J. Am. Med. Assoc. 315 (11), 1141–1148, 2016 Mar 15.

Cremers, H.R., Vager, T.D., Yarkoni, T., 2017. The relation between statistical power and inference in fMRI. PLoS One 12, e01284923. https://doi.org/10.1371/journal.pone.0184923.

David, S.P., Ware, J.J., Chu, I.M., Loftus, P.D., Fusar-Poli, P., Radua, J., Munafò, M.R., Ioannidis, J.P., 2013. Potential reporting bias in fMRI studies of the brain. PLoS One 8 (7), e70104. https://doi.org/10.1371/journal.pone.0070104. Jul 25.

Desmond, Glover, 2012. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118, 115–128.

Durnez, J., Degryse, J., Moerkere, B., et al., 2016. Power and sample size calculations for fMRI studies based on the prevalence of active peaks. Preprint at bioRxiv. https://doi.org/10.1101/049429, 2016.

Faul, F., Erdfelder, E., Lang, A., et al., 2007. *Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175–191. https://doi.org/10.3758/BF03193146.

Friston, K.J., 1999. How many subjects constitute a study? Neuroimage 10, 1–5.

Geuter, S., Qi, G., Welsh, R.C., Wager, T.D., Lindquist, M.A., 2018. Effect size and power in fMRI group analysis. bioRxiv, 295048. https://doi.org/10.1101/295048.

Gorgolewski, K.J., Auer, T., Calhoun, V.D., et al., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 3, 160044. https://doi.org/10.1038/sdata.2016.44.

Guo, Q., Parlar, N., Truong, W., Hall, W., Thabane, L., McKinnon, M., Goeree, R., Pullenayegum, E., 2014. The reporting of observational clinical functional magnetic resonance imaging studies: a systematic review. PloS One 9, e94412.

Hardwicke, T., Ioannidis, J., 2018. Mapping the universe of registered reports. Nat. Hum. Behav. 2, 793–796.

Hayasaka, S., Peiffer, A.M., Hugenschmidt, C.E., Laurienti, P.J., 2007. Power and sample size calculation for neuroimaging studies by non-central random field theory. Neuroimage 37, 721–730.

Ingre, M., 2013. Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: comment on Friston (2012). Neuroimage 81, 496–498.

Ioannidis, J.P.A., 2005a. Why most published research findings are false. PLoS Med. 2, e124.

Ioannidis, J.P., 2005b. Contradicted and initially stronger effects in highly cited clinical research. J. Am. Med. Assoc. 294 (2), 218–228. Jul 13.

Ioannidis, J.P., 2007. Molecular evidence-based medicine: evolution and integration of information in the genomic era. Eur. J. Clin. Invest. May 37 (5), 340–349.

Ioannidis, J.P.A., 2008. Why most discovered true associations are inflated. Epidemiology 19, 640–648.

Ioannidis, J.P.A., 2010. Meta-research: the art of getting it wrong. Res. Synth. Methods 1, 169–184.

Ioannidis, J.P., 2012. Why science is not necessarily self-correcting. Perspect. Psychol. Sci. 7 (6), 645–654. https://doi.org/10.1177/1745691612464056, 2012 Nov.

Ioannidis, J.P.A., 2019. Publishing research with P-values: prescribe more stringent statistical significance or proscribe statistical significance? Eur. Heart J. 40 (31), 2553–2554, 2019 Aug 14.

Lakens, D., Adolfi, F.G., Albers, C.J., et al., 2018. Justify your alpha. Nat. Hum. Behav. 2, 168–171. https://doi.org/10.1038/s41562-018-0311-x.

Lindquist, M.A., Caffo, B., Crainiceanu, C., 2013. Ironing our the statistical wrinkles in the "ten ironic rules". Neuroimage 81, 499–502.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon statistical significance. Am. Statistician 73 (Suppl. 1), 235–245.

Mumford, J.A., 2012. A power calculation guide for fMRI studies. Scan 7, 738–742.

Mumford, J.A., Nichols, T.E., 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261–268.

Munafo, M.R., Nosek, B.A., Bishop, D.V.M., et al., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1, 0021.

Murphy, K., Garavan, H., 2004. An empirical investigation into the number of subjects required for an event-related fMRI study. Neuroimage 22, 879–885.

Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. London, Ser. A 231, 289–337.

Nichols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D.C., White, T., Yeo, B.T.T., 2016. Best practices in data analysis and sharing in neuroimaging using MRI. bioRxiv. https://doi.org/10.1101/054262.

Nichols, T.E., Das, S., Eickhoff, S.B., et al., 2017. Best practices in data analysis and sharing in neuroimaging using MRI. Nat. Neurosci. 20, 299–303.

Petersson, K.M., Nichols, T.E., Poline, J.B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. Phil. Trans. Roy. Soc. Lond. 354, 1261–1281.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafo, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Neurosci. 18, 115–126.

Powell, Michael, Hosseini, Mahan, Collins, John, Callahan-Flintoft, Chloe, Jones, William, Bowman, Howard, Wyble, Brad, 2020. I tried a bunch of things: the unexpected dangers of overfitting. bioRxiv. https://doi.org/10.1101/078816, 078816.

Sedlmeyer, P., Gigerenzer, G., 1989. Do studies of statistical power have an effect on the power of the studies? Psychol. Bull. 105, 309–316.

Simmons, J., Nelson, L., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. Psychol. Sci. 22, 1359–1366.

Smith, P.L., Little, D.R., 2018. Small is beautiful: in defence of the small N design. Psychonomic Bull. Rev. 25, 2083–2101.

Suckling, J., Henty, J., Ecker, C., et al., 2014. Are power calculations useful? A multicentre neuroimaging study. Hum. Brain Mapp. 35, 3569–3577.

Szűcs, D., 2016. A tutorial on hunting statistical significance by chasing N. Front. Psychol. 7, 1444.

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

Szűcs, D., Ioannidis, J.P.A., 2017b. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 15 (3), e2000797, 2 March 2017.

Thompson, William Hedley, Wright, Jessey, Bissett, Patrick G., Poldrack, Russell A., 2019. Dataset Decay: the problem of sequential analyses on open datasets. bioRxiv. https://doi.org/10.1101/801696, 801696.

Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Smaller sample size reduce the replicability of task-based fMRI studies. Commun. Biol. 1, 62, 2018.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p<0.05". Am. Statistician 73 (Suppl. 1), 1–19.

Yarkoni, T., 2009. Big correlations in little studies: inflated fMRI correlations reflect low statistical power—commentary on Vul et al. Perspect. Psychol. Sci. 4, 294–298.

Zandbelt, B.B., Gladwin, T.E., Raemaekers, M., Van Buuren, M., Neggers, S.F., Kahn, R.S., Ramsey, N.F., et al., 2008. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. Neuroimage 42 (1), 196–206.