Review paper: reporting practices for task fMRI studies

Freya Acar [1]
Camille Maumet [2]
Talia Heuten [1]
Maya Vervoort [1]
Han Bossier [1]
Ruth Seurinck [1]
Beatrijs Moerkerke [1]

[1] Faculty of Psychology and Educational Sciences, Ghent University, Belgium
[2] Inria, Univ Rennes, CNRS, Inserm, IRISA UMR 6074, Empenn ERL U 1228, F-35000 Rennes, France

Abstract

What are the standards for the reporting methods and results of fMRI studies, and how have they evolved over the years? To answer this question we reviewed 160 papers published between 2004 and 2019. Reporting styles for methods and results of fMRI studies can differ greatly between published studies. However, adequate reporting is essential for the comprehension, replication and reuse of the study (for instance in a meta-analysis). To aid authors in reporting the methods and results of their task-based fMRI study the COBIDAS report was published in 2016, which provides researchers with clear guidelines on how to report the design, acquisition, preprocessing, statistical analysis and results (including data sharing) of fMRI studies (Nichols, et al., 2016). In the past reviews have been published that evaluate how fMRI methods are reported based on the 2008 guidelines, but they did not focus on how task based fMRI results are reported. This review updates reporting practices of fMRI methods, and adds an extra focus on how fMRI results are reported. We discuss reporting practices about the design stage, specific participant characteristics, scanner characteristics, data processing methods, data analysis methods and reported results.

Key words: fMRI, neuroimaging, reporting method, meta-analysis, CBMA, IBMA

Introduction

The goal of an fMRI study is to localize neural activation during specific tasks or paradigms by detecting the area in the brain that receives and uses the oxygen in the blood (Logothetis, 2008). Compared to other neuroimaging methods, fMRI possesses great spatial precision while maintaining adequate temporal precision (Mehta & Parasuraman, 2013).

Since it was first employed in the early 1990's and after gaining popularity in the early 2000's, there has been a great evolution in the scanners that are employed, the analysis software that is used, the study paradigms and, most importantly, in the reporting standards. On the one hand, there are the evident steps of an fMRI study that need to be reported such as study design and participant characteristics, but on the other hand, there is also a multitude of specific details such as scanner characteristics and data processing settings.

An accurate and complete description of empirical research is the cornerstone of good scientific practice. In a first step, research needs to be thoroughly evaluated by other experts in the field as in the process of peer review. Secondly, reporting should enable replication by potentially other researchers to gain insight into the validity of results, within the same or a different context. Third, meta-analytic approaches are commonly used to explain variation in observed study results and to combine results over different studies. None of these three steps (evaluation, replication and summarization) is possible without complete coverage of the empirical process.

In 2008 Poldrack, et al. published guidelines for reporting an fMRI study. In this publication they discuss the necessary information that should be reported on participants and the task they were required to perform and point out the difference between standard space and templates. With respect to statistical analyses, they instruct to specify how regions of interest were determined, which software package was used and which user-specified characteristics were changed in the software package. They discuss how a description is needed on how group effects and effects in individual participants are analysed and what statistical tests and thresholds are used for inference. The paper provides a structured overview of all characteristics to be reported and is a very useful guideline for reporting design and statistical analyses. However, it does not provide guidelines on how to report the results of fMRI studies.

An update to these guidelines was formulated in 2017 by Poldrack, et al. In this paper the effect of using different analysis "pipelines" (or procedures) is demonstrated and discussed, highlighting the importance of proper reporting practices. There is also a small section included on how results can be shared, yet it is not very extensive. Finally, he COBIDAS report (Nichols, et al., 2016) is the most comprehensive document that provides readers with thorough guidelines on how to report the design, acquisition, pre-processing, statistical analysis and results (including data sharing) of fMRI studies. Based on these guidelines, authors can report the methods and the results of their task-based fMRI study

This leaves the question how authors report the methods and results of fMRI studies in the past and now. Up to now two papers have been published that evaluate reporting practices in fMRI studies (Carp, 2012; Guo, et al., 2014), based on the guidelines formulated by (Poldrack, et al., 2008). This includes experimental design, data acquisition, data processing, statistical modelling and visualization. This publication focuses mainly on the methods section, including the analysis pipeline, and does not discuss any practices on how the results of task-based fMRI studies are reported.

Due to the specific format of fMRI study results, results are difficult to report and publish in full. fMRI study results consist of brain images with often over 200.000 data points. Printing approximately

200.000 data points on paper is tedious, and unnecessary as no one will be able to use the results in this format. Publishing the results digitally requires space, as the results of an fMRI study can take up several gigabytes. Even though online repositories are available that allow for the publication of fMRI results (i.e. raw data on OpenNeuro (https://openneuro.org/, (Markiewicz, 2021)) or statistical maps on NeuroVault (Gorgolewski KJ, 2015)), fMRI studies still often merely report the most significant results of a study. These are the locations (sometimes accompanied by a test statistic) in the brain with the highest peak. The degree of information that is made available however determines how results of a single study can be used in a broader meta-analytic context.

Hence, fMRI studies face reporting challenges at various levels. In this paper we focus on reporting practices for materials, methods and results of fMRI studies. Since the COBIDAS report was published (Nichols, et al., 2016), no reviews have been published that evaluate the impact of these guidelines on reporting practices. Furthermore, no reviews have been published that describe reporting practices for fMRI results. It is generally accepted that merely the location of local maxima are reported in fMRI studies and that statistical maps are often overlooked, but this has never been shown in a peer-reviewed publication.

To get a clear overview of reporting practices for materials, methods and results in the field of fMRI we briefly describe the different steps of an fMRI study and select characteristics that are essential and require reporting. We also list the possible ways in which fMRI studies can be reported and evaluate papers that were published between 2003 and 2020 for these characteristics. In this way, we provide an update on the review by Carp (2012) and Guo et al. (2014)after the publication of the 2016 COBIDAS report and with due attention for the reporting of results. This paper is the first to evaluate in detail reporting practices for fMRI studies, how they evolved, and where the greatest opportunity for improvement lies.

In the methods section we explain how the review was performed. We elaborate on the selection criteria. We detail which characteristics were coded and why they were chosen. In the results section we present our findings concerning the previously chosen characteristics. Finally, in the discussion, we view the results in their context and compare them to the results obtained by (Carp, 2012; Guo, et al., 2014).

Methods

Article selection

We provide an overview of reporting practices in fMRI studies over the years. To select candidate papers, we used the same criteria as (Carp, 2012). A PubMed search was conducted on 31/10/2020 to identify eligible articles. The search criteria were: the abstract mentioning ``fMRI", ``functional MRI" or ``functional magnetic resonance imaging" and excluding studies mentioning ``resting state" or ``connectivity". Solely articles published in English between 2004 and 2019 were retained. Only articles for which the full text was accessible were considered. The flowchart can be seen in

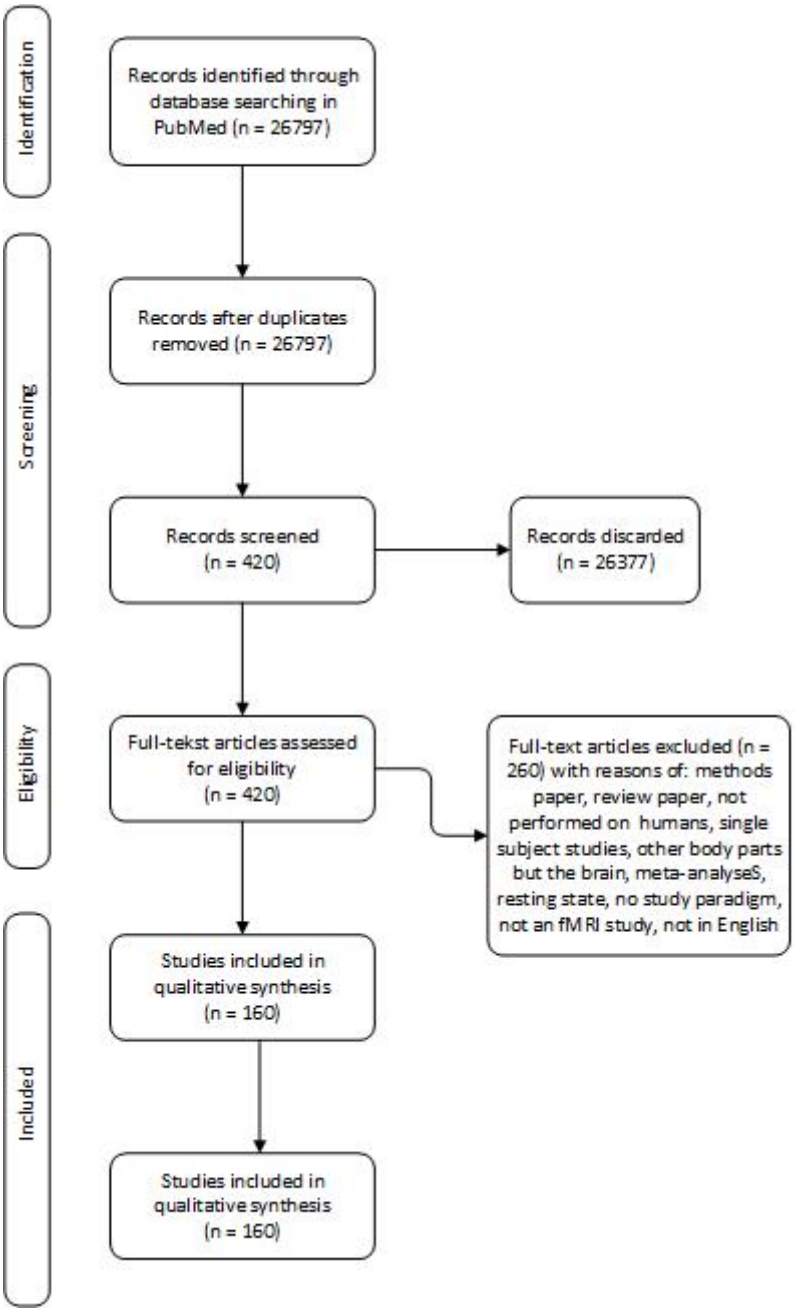

Figure 1. Flowchart exclusion process..

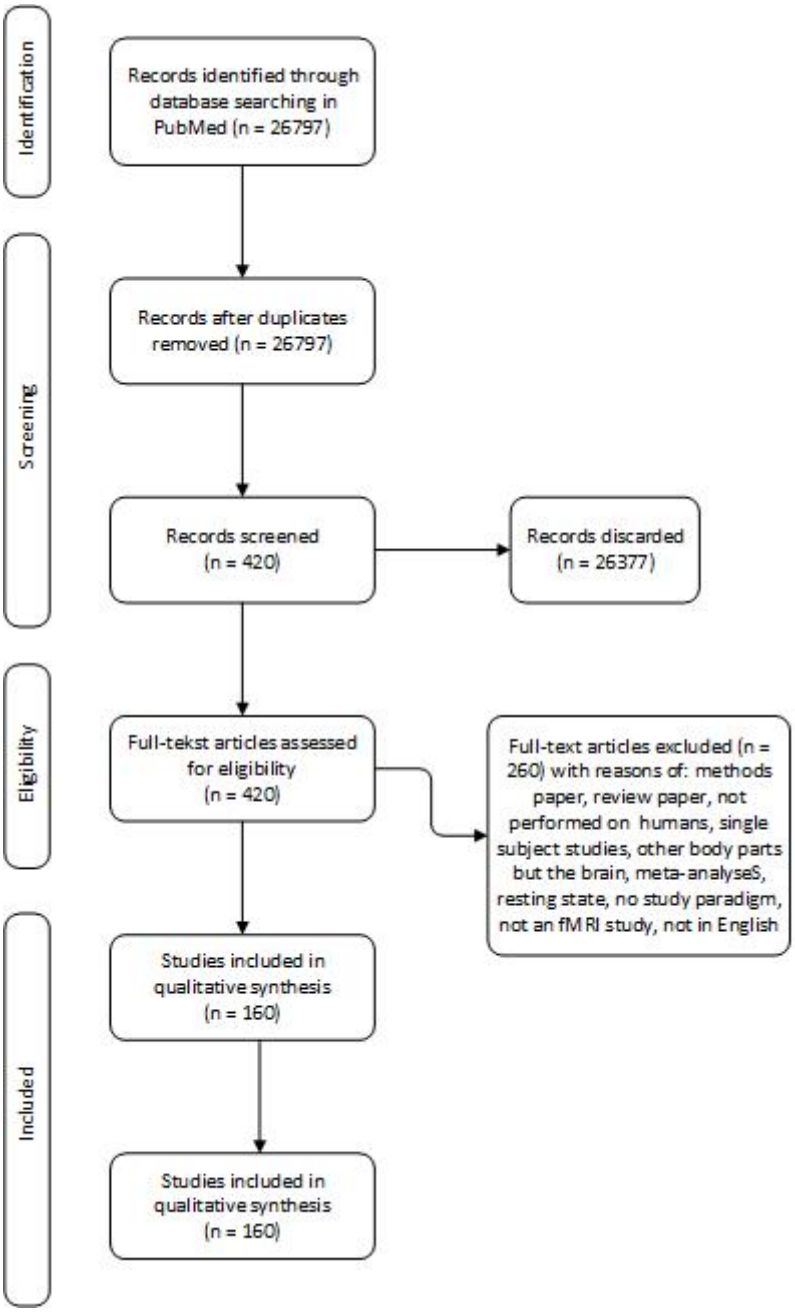

Figure 1. Flowchart exclusion process.

Based on these criteria, 26797 articles were selected. The papers were grouped by year of publication, and randomly ordered. One by one the papers were read and entered in the review or excluded, until 10 papers for every year of publication were selected for the review. In this review we do not include other reviews or meta-analysis, and no method papers, but only single studies. We focus on task fMRI studies and exclude all resting state fMRI studies as these differ greatly in methods. Furthermore we exclude all studies with only a single subject or that are performed with animals and not with humans. Finally we only include studies that focussed on the brain in the MRI scanner, and excluded studies that investigated other parts of the body. In Table 1. Reason and number of papers that are excluded from the review and the number of exclusions per criterium are listed.

Table 1. Reason and number of papers that are excluded from the review

REASON FOR EXCLUSION	NUMBER OF EXCLUDED PAPERS FROM REVIEW BECAUSE OF THIS REASON
METHODS PAPER	127
REVIEW PAPER	60
NOT HUMAN	36
SINGLE SUBJECT	15
NOT BRAIN	8
META-ANALYSIS	6
RESTING STATE	3
NOT A STUDY	2
NOT FMRI	2
NOT IN ENGLISH	1

The papers that met inclusion criteria were then carefully and independently read by two researchers to ensure the validity of the results. Both researchers read the papers to assess the reporting practices of the papers and wrote down their results independently. If the researchers wrote down different outcomes the first author of this manuscript checked the research paper again to understand what caused the difference and write down the correct result. The final results were written down in an excel file.

Study characteristics

We aim to provide an overview of study characteristics, and how they have evolved in the last 15 years. To structure the results of this overview, these characteristics were divided into 7 "themes" such as "study design" and "scanner characteristics". Here we will describe the themes and how we registered results for every theme.

Study identification

The first batch of information we gathered from every paper refers to the identification of the study. This information is written down for every paper that was read to check the exclusion criteria. The next information parts are only registered for studies that are included in the review. We used the PubMed database to search for relevant publications. Every publication on PubMed has a unique ID. Next to the title and the year of publication, the PubMed ID was saved. Not all publications that were found during the search are included in the review. If a publication was excluded, for instance because it was a study performed on animals, a meta-analysis, or a single subject study, the reason was written down.

- **Identification**: everything that is necessary for the identification of the study
    - PubMed ID (number)
    - Year of publication (number)
    - Title of the publication (text)
    - If the paper was excluded, relevant exclusion criterion

Study design

The first step in the execution of an fMRI study is the design. What do we want to investigate? Which paradigm will we use? What is the timing of the stimuli?

The first characteristic we investigate is whether the paradigm of the study is thoroughly described or not. Have the authors described the setup of the study or not? This includes the different conditions of the study paradigm that are presented to the participants and their relevant timing.

The second characteristic is the type of design. To identify which brain regions are active during a certain task the difference in brain activation is discerned during two possible states: performing tasks or not. There are two common possibilities in which the task can be presented. It can either be done through a block design, where task and non-task "blocks" alternate. In this case the period of time where the task is activated is compared to the period of time of the non-task block. The other option is an event-related design, where the majority of the time no task is presented, and every now and then a short "event" of the task is presented to the participant. A combination of the two designs is also possible.

The third parameter we consider is design optimisation (Kao, Temkit, & Wong, 2014). Design optimisation involves constructing an optimal design matrix (i.e. relative timing and frequency of presenting stimuli) to obtain statistically efficient estimators. One aspect of design optimisation that is frequently used is inter-trial jittering. An MRI scanner cannot scan the entire brain in one go, the brain is scanned slice by slice. If for instance the timing of the scanner is synchronized with the timing of the stimuli, it is possible that by chance the HRF peak and corresponding activated area is missed every time. It is even possible that another, irrelevant, area of the brain is scanned every time the task is presented. To prevent this problem it is wise to optimize the study design, for instance through jittering. When a design is jittered the interval between two tasks is of varying length. To evaluate whether the design of a study was optimized or not we looked for the words "design optimisation" and "jitter" in the text.

The fourth characteristic verifies whether more than one experiment was included in the paper, this can either be another fMRI experiment, or a behavioural experiment. If more than one fMRI experiment is included, the first experiment is used for further analysis.

Finally, the fifth characteristic indicates whether the fMRI experiment was the main focus of the research paper, and whether the univariate approach was used for the analysis. This was done to ensure that the papers included in this review need to report the same information in their methods and results section. Other methods and approaches require other reporting styles. If the fMRI experiment is not the main focus, this might influence the detail with which it is reported. We further indicate the type of analysis (mass-univariate versus multivariate) since in this paper, we focus on characteristics of univariate analyses.

- **Design**: specific design characteristics of the study
    - Is the design described? (0 = no, 1 = yes)
    - Type of design (list: block or event-related)
    - Was the design optimized? For instance, were the stimuli jittered? (0 = no, 1 = yes)
    - Was more than one experiment included in the paper? (0 = no, 1 = yes)
    - Was the fMRI experiment and univariate approach the main focus of the paper? (0 = no, 1 = yes)

Participant characteristics

The third theme is the participant characteristics. fMRI studies may be underpowered because of the large number of statistical tests that are typically performed in combination with a modest sample size. A larger set of participants increases power, but fMRI experiments and adding more participants

to them are quite costly. Therefore, we want to verify the average number of participants in fMRI experiments, and whether this number evolved over time.

The first characteristic is the number of participants for every fMRI experiment.

The sample is typically assumed to be representative for the entire population, or part of the population. In the past few years researchers have realized that bias caused by the way samples are built poses a risk for the validity of the study outcome. One possible cause of bias is a gender imbalance between the participants of the study (Larrazabal, 2020). Therefore we verify gender balance in this parameter. In characteristics 2, 3 and 4 we respectively consider the number of female and male participants per experiment, and the ratio female and male participants.

By selecting a subset of participants (e.g. removing outliers), results may change drastically If this goes undocumented or poor argumentation for removing data is given, one should be aware about the possible introduction of bias. On the other hand, sometimes it is necessary to remove participants from the results, if they for instance moved too much during the scanning. To ensure that participants were legitimately removed from the study in a transparent manner, we keep track of whether the exclusion criteria for participants were described in the paper or not. Here we focus on participants that were removed from analysis after they took part in the fMRI experiment and the data had been administered. We do not take into account participants that are excluded before they take part in the experiment (for example because of having tattoos).

Finally, we verified whether a power analysis was done to determine the required sample size, or whether power of the study was mentioned. Power analyses for fMRI studies are complex to perform. Nevertheless, because of the large number of statistical tests that are being performed the sample size needs to be adequate to attain sufficient power to detect true activation.

- **Participant characteristics**
    - Number of participants (number)
    - Number of female participants (number)
    - Number of male participants (number)
    - Ratio of female and male participants (number)
    - Are the exclusion criteria for participants described in the paper? (0 = no, 1 = yes)
    - Was a power analysis performed to determine the required sample size?

Scanner characteristics

The fMRI scanner is the machine in which the participants take place during a task and that scans their brain. Since they were first developed scanners have evolved. Therefore, the fourth section writes down scanner characteristics.

The first characteristic is the field strength of the scanner. This indicates the resolution with which the brain could be scanned. When the scanner has a higher resolution, for instance 7T, smaller parts of the brain can be inspected. Consequently, because the brain is divided into more slices, it takes longer for the entire brain to be scanned, increasing the chance that an activated area is missed due to timing.

The second characteristic is whether the whole brain is scanned or not. One possible solution to not miss the activated area is to only scan the part of the brain where activation can be expected. This might be a solution to overcome a problem for a single study, but it impedes the study to be entered into a meta-analysis.

The third characteristic considers voxel resolution. During scanning the brain is divided into small cubes, voxels, where the presence of activation is evaluated. The size of these voxels is determined by

the resolution of the scanner, the number of slices that are administered and more. A lot of variation is visible in voxel size.

Finally, the dimensions of the scanned image vary. This is the number of voxels in every dimension that are evaluated. This characteristic also depends on the scanner resolution, the number of slices etc. The dimensions of the scanned image, in number of voxels in every dimension, are written down in characteristic 5.

- **Scanner characteristics**
    - Field strength of the scanner in Tesla (list: 1.5, 3, 4, 7)
    - Was the whole brain scanned? (0 = no, 1 = yes)
    - Voxel resolution (number x number x number in mm)
    - Dimensions of the scanned image (number x number x number in voxels)

Data processing

Once the data has been gathered, it needs to be processed before analysis. Individual data from the participants needs to be prepared for analyses across participants. Different software packages are designed to pre-process data but may differ in how this is done. A first characteristic of interest is therefore the software that is used. The most often employed software packages are FMRIB Software Library (FSL, homepage: http://www.fmrib.ox.ac.uk /fsl), Statistical Parametric Mapping (SPM, homepage: http://www.fil.ion.ucl.ac.uk/spm) and Analysis of Functional NeuroImages (AFNI, homepage http://afni.nimh.nih.gov). All software packages can be used for the pre-processing and analysis of fMRI studies, but it has been shown that if the same study is analysed by all three packages the results will differ significantly across these software packages (Bowring, Maumet, & Nichols, 2019).

The second characteristic is a score for how well the pre-processing steps are described. We identify three pre-processing steps and for every step give a score on how well it was described. 0 if the step was not mentioned, 1 if it was mentioned and 2 if it was clearly described. The final score ranges from 0 (the pre-processing was not mentioned at all) to 6 (every step was clearly described).

The first pre-processing step we consider is motion correction, i.e. the correction of movement of participants in the scanner. Taking into account that one voxel is typically 3x3x3mm, a movement of 3mm already places detected activation in a different voxel. The degree of motion of a participant can be estimated and, where possible, corrected. Nevertheless, if movements of a participant are too large or too frequent, some trials or the entire run needs to be omitted from the study results. We view the motion correction step as clearly described when the reference image is reported. It is considered mentioned when something similar to “motion correction was performed” was written down in the article but was not further described in more detail.

Motion parameters can be included in the design matrix. By doing this the effect of motion on the results is reduced. We verified whether the inclusion of motion parameters in the design matrix was mentioned or not.

Just as every person is different, their brains differ as well. Generally, structures are similar, but to verify possible group effects their brains need to be mapped onto each other. Usually, a standard coordinate space is used and the brain of every individual participant is mapped onto that coordinate space, reshaping their brains. The brains of some participants are enlarged during this process, some brains are made smaller. Whether this process is clearly described (the standard space and the transformation method are reported), mentioned, or neither is written down in the second part of the second characteristic of this section. Furthermore, there are several options for the coordinate space

onto which the participants' brains are mapped. Which coordinate space is used is the third characteristic of this section.

The last pre-processing method to correct for participant movement and brain differences is spatial smoothing and temporal filtering. Using spatial smoothing,, the BOLD signal is smoothed out over neighbouring voxels. Technically, this implies that the value in each voxel is replaced by a weighted average of its own value and neighbouring voxels (the further away, the smaller the weight) to obtain a spatially smooth image. Several data-analytical techniques rely on the assumption of smoothness. After smoothing, activation is no longer limited to one voxel and there is a higher chance of detecting a pattern across participants, but some spatial precision is lost in the process. The size at the full-width half maximum (FWHM) of the smoothing kernel is noted down as the fifth characteristic of this section. If the process of smoothing and the kernel width are mentioned this step is considered "clearly described".

In addition, temporal filtering can be applied by setting a highpass filter to remove low frequency signal from heart rate, breathing, etc.

For every step a score is given from 0 to 2, and eventually a score out of 6 is given to every study for how well these pre-processing steps were described.

- **Data processing**
    - Which software was used for data processing? (text)
    - A score for pre-processing steps determined by (number between 0 and 6)
        - Motion correction (0 = not mentioned, 1 = mentioned, 2 = clearly described)
        - Motion regression (0 = not mentioned, 1 = described)
        - Registration (0 = not mentioned, 1 = mentioned, 2 = clearly described)
        - Spatial filtering (0 = not mentioned, 1 = mentioned, 2 = clearly described)
        - Temporal filtering (0 = not mentioned, 1 = mentioned, 2 = clearly described)
    - Which coordinate space was used (text)
    - FWHM of the smoothing kernel (number)

Data analysis

Once the data has been pre-processed and is qualitatively in its most pure form, it is time to start with the data analysis. For the data analysis two elements are required: the design of the study and the hemo-dynamic response function (HRF). Based on the design of the study a design matrix can be constructed that indicates which events were present at every timepoint. The HRF, which represents the specific pattern of the response of oxygenated blood flow, is estimated based on the design matrix and the observed patterns. The HRF can differ across participants and is essential for the data analysis. Therefore, the first characteristic is whether the HRF is mentioned, and the second characteristic which model was used for the HRF.

During the analysis, typically the contrast between two (or more) conditions is evaluated (e.g. task present and task not present). To be able to interpret study results it is essential to define the contrast unequivocally. Therefore, the third characteristic is whether the contrast is clearly described. During the data analysis the states get a value (e.g. +1 if the task is present and -1 if the task is not present). This value can arbitrarily be chosen, and even be weighted (e.g. +1/4 and +1/6 in two situations where the task is present and -1/2 and -1/2 in two situations when the task is not present). As a general rule of thumb, to avoid scaling problems, Nichols advises to ensure that the sum of all positive contrast elements is equal to 1, and the sum of all negative contrast elements is equal to -1 (Nichols, 2012). Whether or not the scaling of the contrast is described is noted in characteristic 4.

The scaling of the predictors ensures that, at the first level (participant level) analysis, the regression coefficients have the same units as the data (Nichols, 2012). To correctly scale the predictors, the baseline to peak magnitude should be 1.0, which implies that if the beta coefficient changes with one unit, the predicted BOLD effect will change with one unit as well. This eases interpretation of the effect of predictors but there exist substantial differences in how scaling is implemented between different software packages (Bowring, Maumet, & Nichols, 2019) Whether the predictor scaling is mentioned or not in the paper is written down in characteristic 5.

First, the analyses are performed at participant level. The seventh characteristic notes the statistical model and estimation method that is used at the first level, the participant level. Then the results of all participants are analysed together on the second level, the group level. Alternatively, the second level consists of pooling runs within participants before proceeding to the group analysis. The eight characteristic is the statistical model and estimation method used for the group analysis. The sixth characteristic notes whether the interpretation of the contrast at group level is clear or not. This is closely related to the third characteristic that notes whether the contrast is clearly described or not.

After the group analyses have been performed the locations in which the activation is larger than can be expected by chance is determined. This is done through an inference method. Inference can be done at voxel-level, where for every voxel it is determined whether its activation is larger than can be expected by chance or not. However, it can be reasoned that activation will most likely be spread across several voxels that lie next to each other. Cluster-wise inference methods take the spatial component into account. In characteristic 9 we write down the inference method (voxel-wise or cluster-level). In characteristic 10 the cluster-forming threshold is noted in case of a cluster-level inference method. Then in characteristic 11 the inference threshold is written down and in characteristic 12 inference method. Finally, characteristic 13 is reserved to describe remarks about the thresholding method, for instance if a region of interest analysis (ROI) was performed (not the whole brain but only a part, a region of interest, is analysed to reduce the number of statistical tests and hence increase power).

- **Data analysis**
    - Is the HRF model mentioned? (0 = no, 1 = yes)
    - Which model was used for the HRF (text)
    - Was the scaling of the contrast described? (0 = no, 1 = yes)
    - Was the scaling of the predictors reported? (0 = no, 1 = yes)
    - Is the interpretation of the contrast at group level clear? (0 = no, 1 = yes)
    - What is the statistical model and estimation method used at the first level? (text)
    - What is the statistical model and estimation method used for the group analysis? (text)
    - Inference method (list: voxel-wise, cluster-level)
    - In case of topological inference, what was the cluster-forming threshold? (text)
    - Inference threshold (number)
    - Inference method (text)
    - Remarks about the thresholding method (text, e.g. if ROI analyses were performed)

Reporting characteristics

Once the results are finalized they need to be reported. Whether the results were analysed and reported on whole brain level or not was chosen as the first reporting characteristic. fMRI study results consist of maps of the entire brain. One map contains the contrast results, the difference in activation in the brain between the situation where the task is being performed and when it is not, summarized over trials and participants. In characteristic 6 we note whether this map is shared. A second map

contains the standard errors. In characteristic 7 it is noted whether the map with standard errors is shared or not.

There are other ways to summarize these maps into one statistic. The first option is through standardized effect sizes, this is noted in characteristic 2. Other options are statistical maps. In characteristic 3 we note whether the statistical maps are shared or not. In characteristic 4 we write down how the maps were shared (in a database, after sending an e-mail to the authors etc.). Characteristic 5 notes which test statistics, such as z-maps, t-maps or p-maps are shared, if in characteristic 3 it is written down that statistical maps were shared.

The most common option to share fMRI study results is by writing down the xyz-coordinates of local maxima, sometimes accompanied by their test statistic. During this process a lot of data and information is lost, but is the most straightforward method of sharing fMRI study results. In characteristic 8 it is noted whether local maxima are shared with a test statistic, and characteristic 9 whether they are shared without a test statistic.

To all authors that did not publish the study results as image maps an e-mail was sent to check whether the data is available somewhere, or if it is available upon request. The results of this e-mail are noted in characteristic 10.

- **Reporting characteristics**
    - Were results analysed and reported on whole brain level? (0 = no, 1 = yes)
    - Are maps with standardized effect sizes available? (0 = no, 1 = yes)
    - Are statistical maps available? (0 or 1)
    - How were the maps shared? (after e-mail, in which database, …) (text)
    - Which test statistics were shared (e.g. z-maps, t-maps, p-maps)?
    - Are contrasts or characteristic maps shared? (0 = no, 1 = yes)
    - Are standard errors shared? (0 = no, 1 = yes)
    - Were the local maxima shared accompanied by a test statistic? (0 = no, 1 = yes)
    - Were the local maxima shared (without a test statistic)? (0 = no, 1 = yes)
    - Was the data available after an e-mail was sent to the authors? (0 = no, 1 = for collaboration, 2 = could be made publicly available, -1 cannot reach author)

Results

The table with results can be viewed in the appendix A: For every characteristic we will first describe how often it was reported adequately (e.g. was the design reported), and continue by looking deeper into what is reported (e.g. which design was used by studies).

Study Design

We see that 137 out of 160 studies describe the study design, 85,63%. We could not see a trend over time, and no statistically significant difference was detected, as the slope is not different from 0 (t = -0.764, p = 0.458). With 6 out of 10 studies that describe the study design, 2008 and 2016 are the years with the smallest number of study design descriptions (see Figure 2. Typically studies).

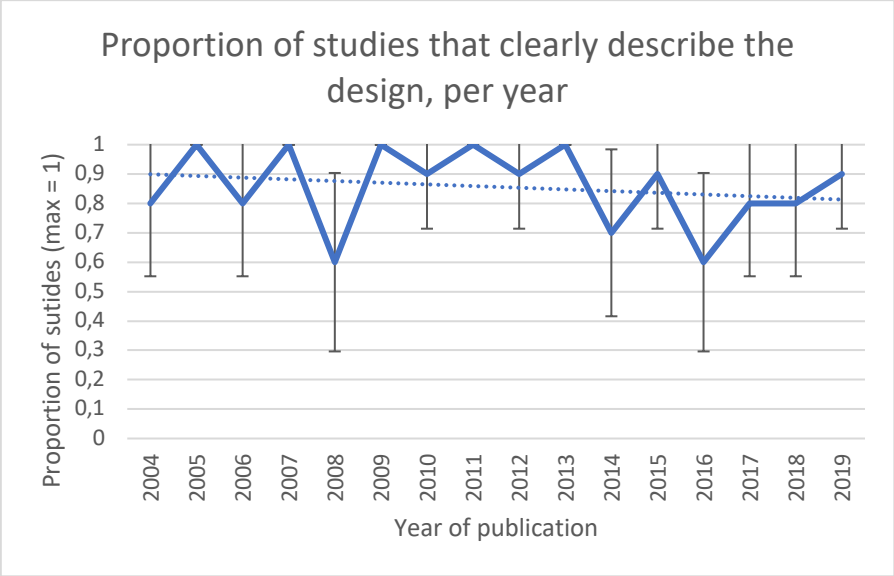

Figure 2. Typically studies clearly describe the study design. There is no clear trend visible over time, but in recent years designs have become more complex, making it harder to describe them in a precise manner.

A block design is the most frequently chosen design, 106 times out of 160 studies. Older studies are more likely to employ a block design. 33 studies mention an event-related design and 4 studies a mixed design. In 2012 6 studies with an event-related design were reported, which is the largest number. In total 14 studies do not clearly mention the type of design, and this occurred most frequently in recent years (see Figure 3. Type of design that was chosen, in function of year of publication.).

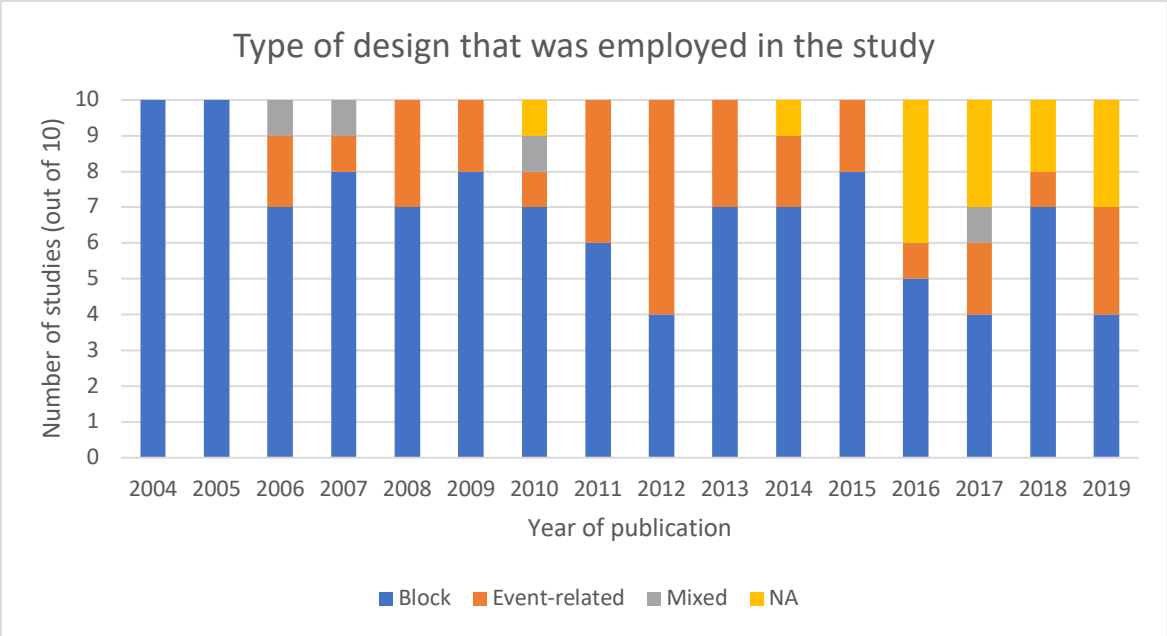

Figure 3. Type of design that was chosen, in function of year of publication. We see that a block design was most popular in the early 2000's, then event-related designs became more common. A mix of block and event-related design is rarely employed. In more recent years we see that the type of design is less often clearly mentioned and that designs become more complex.

15 out of 160 studies mention design optimization and a trend is visible where more recent studies more often employ an optimized design (see Figure 44), however this trend is not statistically significant (t = 1.563, p = 0.14).

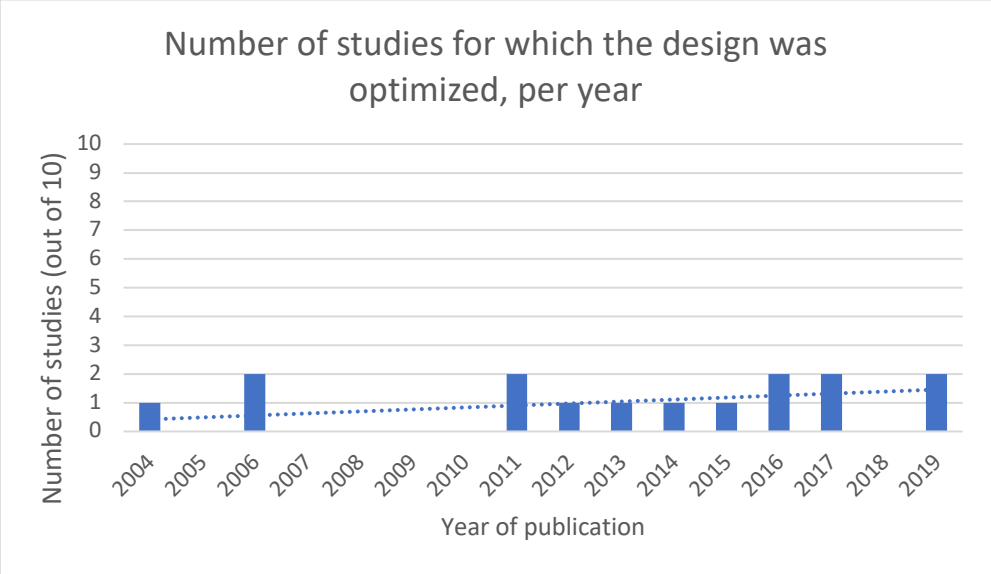

Figure 4. To overcome the possibility of bias caused by the way the brain is scanned the design can be optimized, for instance through jittering. When the design is jittered the scanner scans a different area of the brain during the experimental condition, ensuring that activation is not missed because of the design of the study. We see that design optimization became more common since 2011, but a large majority of studies never reported the use of design optimization.

In 31 out of 160 studies more than one experiment was reported in the study, and in 20 out of those 31 studies the fMRI experiment and a univariate approach was the main focus of the paper.

Participant Characteristics

All studies report the number of participants.

The number of participants ranges from 2 (since single subject studies were excluded) to 283, the average is 34 and the median 26. Over the years a trend is visible where earlier studies employ smaller sample sizes (e.g. an average of 15.7 participants in 2004) and later studies larger sample sizes (e.g. an average of 52.8 participants in 2019, see Figure 55) . This is statistically significant (t = 3.165, p = 0.007).

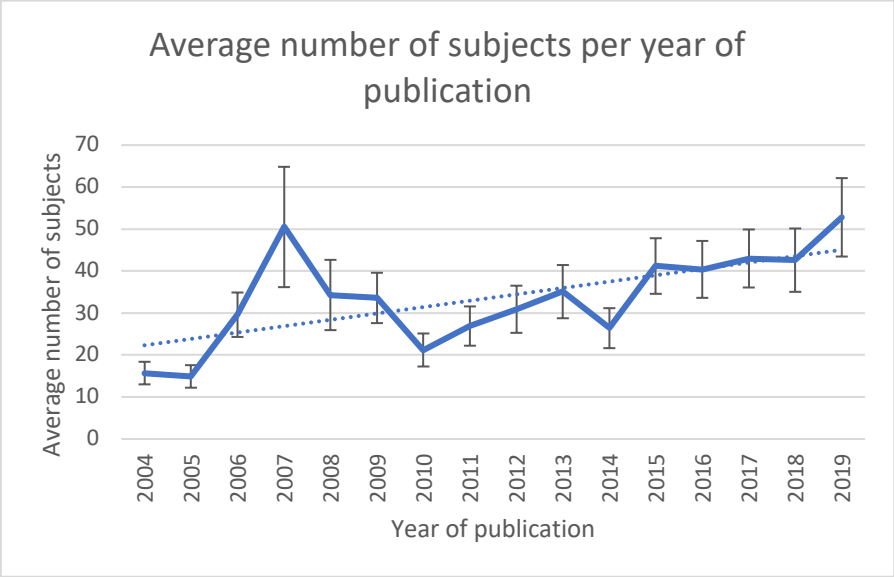

Figure 5. Evolution of the average number of participants. We see that the average number of participant rises over the years, with a peak in 2007 because of 1 study with a large number of participants. In 2004 the average number of participants was only 15, in 2019 this number became almost 4 times lager.

Not all participants that start a study are eventually entered into the results. We registered how often the exclusion criteria were clearly described. A lot of variation is visible (see Figure 6). In 2004 only 3 out of 10 studies clearly described the exclusion criteria, this was 9 out of 10 studies in 2019. The trend of more studies clearly describing the exclusion criteria for participants is not statistically significant (t = 1.634, p = 0.125).

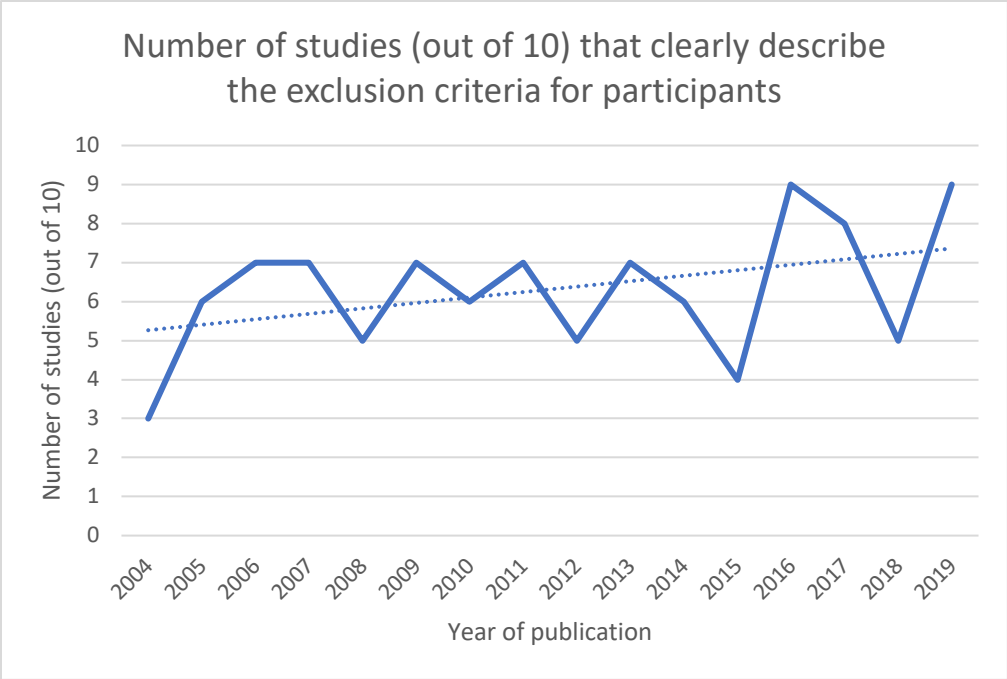

Figure 6. Evolution of the reporting of exclusion criteria for participants. A lot of variation is visible throughout the years, nevertheless a trend is clearly visible where the exclusion criteria for participants are more precisely described in more recent years.

Three studies performed a power analysis to determine sample size before the start of the study, 7 studies adapted their design or reported conducting less analyses to avoid loss of power, 2 studies performed post hoc power analyses and 23 studies indicated that the study might suffer from low power to detect true activation (see Figure 7). One study performed a power analysis for the behavioural experiment, but not for the fMRI experiment. 124 (77.5%) studies did not mention power analysis at all.

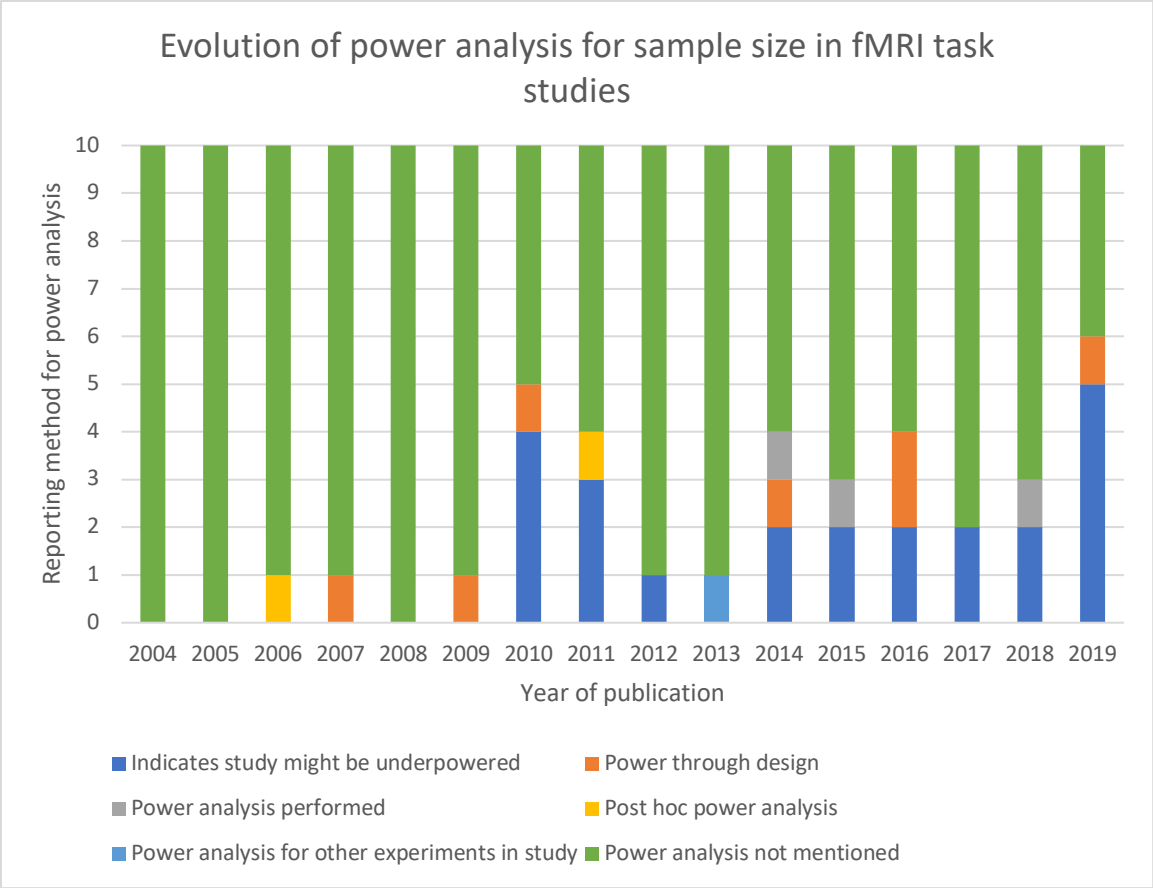

Figure 7. Power analyses are rarely mentioned. Only in 2014, 2015 and 2018 there was one study in our sample that performed a power analysis to determine sample size of the fMRI study. From 2010 on we note studies where it is indicated that the results might be biased because they might be underpowered.

Scanning Characteristics

All studies mentioned scanner field strength. fMRI scanners differ in field strength. Most studies employ 1.5T (70 studies) or 3T (86 studies), while 4T (3 studies) and 7T (1 study) are less frequently employed. Before 2009 most studies employed 1.5T scanners, while after 2009 a switch is visible where 3T scanners are more frequently employed (see Figure 88).

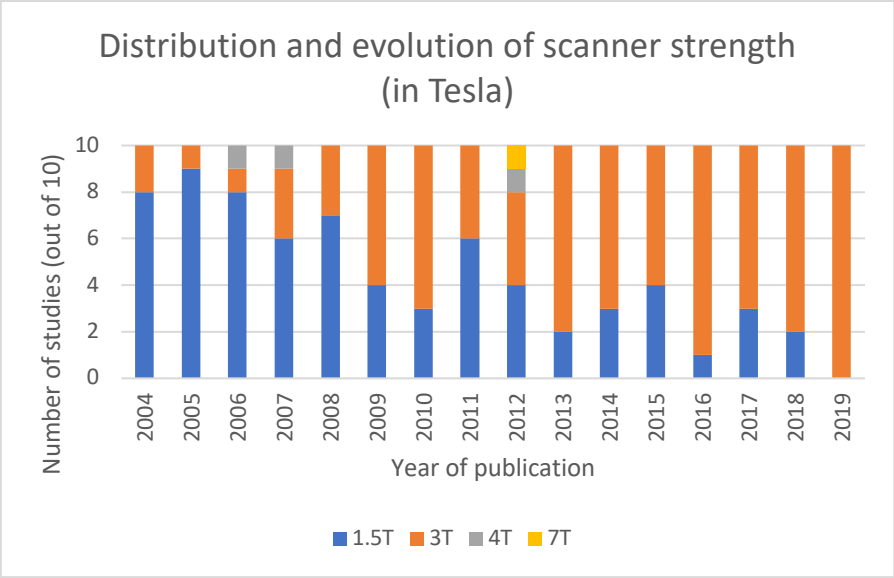

Figure 8. Which scanner strength was used in the studies, and how did this distribution evolve over the years. In every study scanner strength was mentioned.

In 47 out of 160 studies the voxel resolution is not mentioned. Likewise, in 37 out of 160 studies the dimensions of the scanned image are not clearly described. In this case it is most often the number of axial slices is not mentioned.

As shown in Figure 9 all studies scanned the whole brain up until 2009. From 2010 on studies that only scan a region of interest started popping up. In 2014 and 2018 we see the largest number of studies that did not scan the entire brain. This trend of more recent studies less often performing whole-brain analysis is statistically significant ($t = -2.482$, $p = 0.026$)

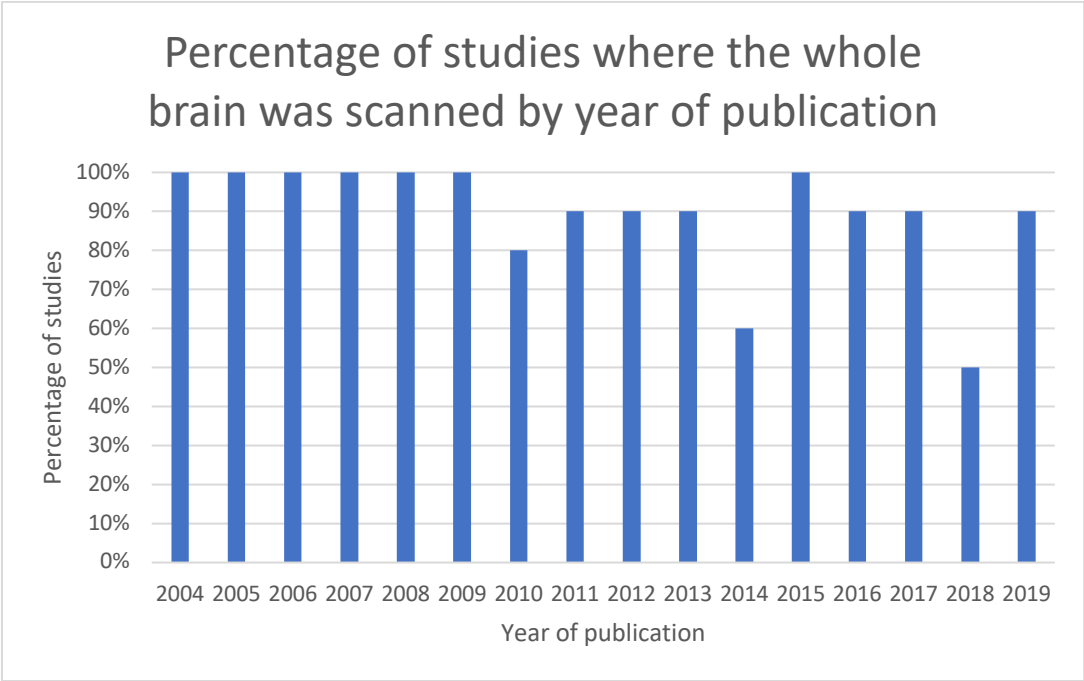

Figure 9. Percentage of studies where the whole brain was scanned and the evolution of this distribution over the years.

Data Processing

Eleven out of 160 studies did not clearly report the data processing package that was used, or employed in-house scripts that were not shared.

Four main data processing packages have been identified: FSL (23 studies), SPM (94 studies), AFNI (15 studies) and BrainVoyager (17 studies), for 11 studies an in-house script was used or the software was unknown (these are notes as "NA" in Figure 10). BrainVoyager was more popular in the earlier studies, SPM has always been the most frequently employed package but the number of studies employing SPM increased over the years (see Figure 10. Software package employed to process the data, and the evolution over the years. It is visible that SPM is the most employed software package. As previously mentioned the chosen software package has a significant effect on the results of the study (Bowring, Maumet, & Nichols, 2019).

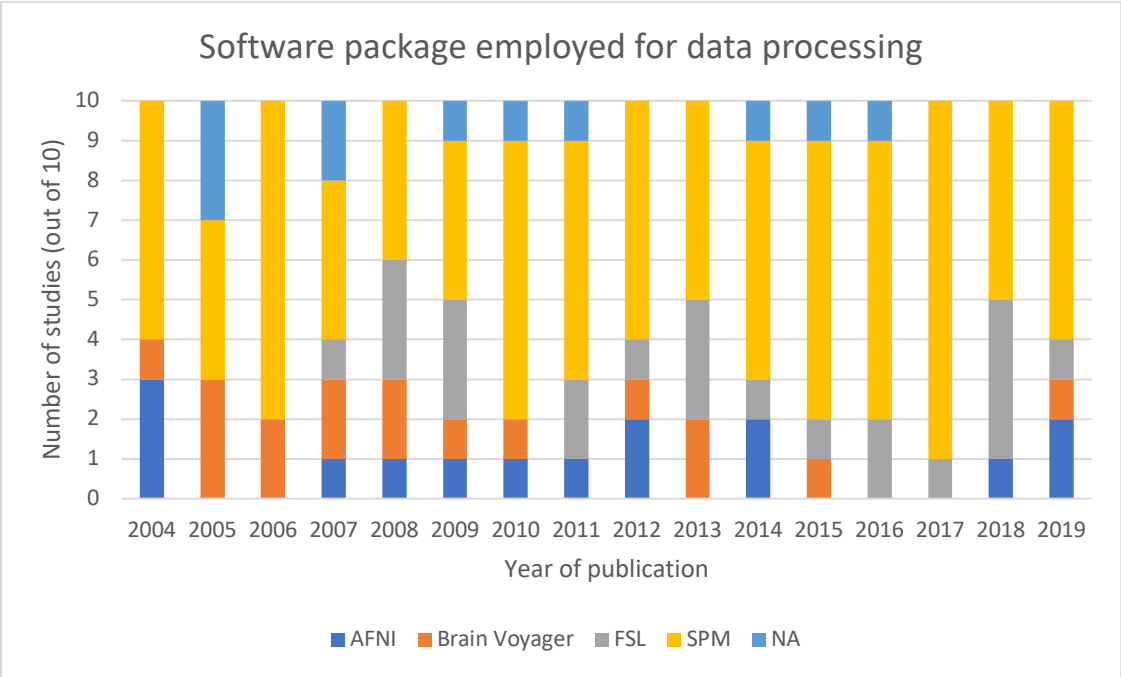

Figure 10. Software package employed to process the data, and the evolution over the years. It is visible that SPM is the most employed software package. "NA" is used for studies that employed an in-house script or where the software was not mentioned.

The reporting of the preprocessing steps was scored by assessing the description of motion correction, registration, spatial filtering and temporal filtering. For all four parameters the study could get a score of 0 if nothing was mentioned, 1 if it was mentioned and 2 if it was clearly described. Therefore the maximum score for preprocessing in total is 8, the minimum score 0. The lowest average was in 2004 (4.3), the highest in 2018 (7.2). The score increases over time. Motion correction is most often clearly described (score of 257 out of 320), followed by spatial filtering (256/320), registration (241/320) and lastly, temporal filtering (162/320). This can be seen in Figure 1111. Furthermore, in Figure 12, we see the evolution of reporting on the inclusion of motion regressors in the design matrix.

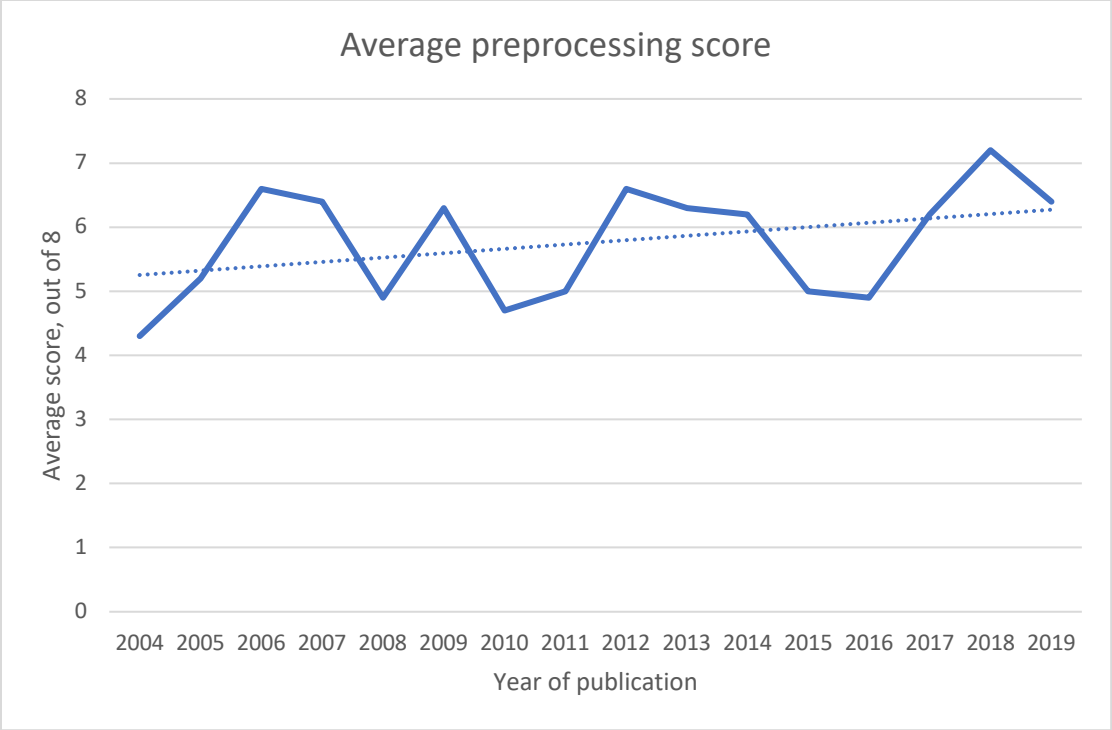

Figure 11. Evolution of scores for how clearly preprocessing steps were described.

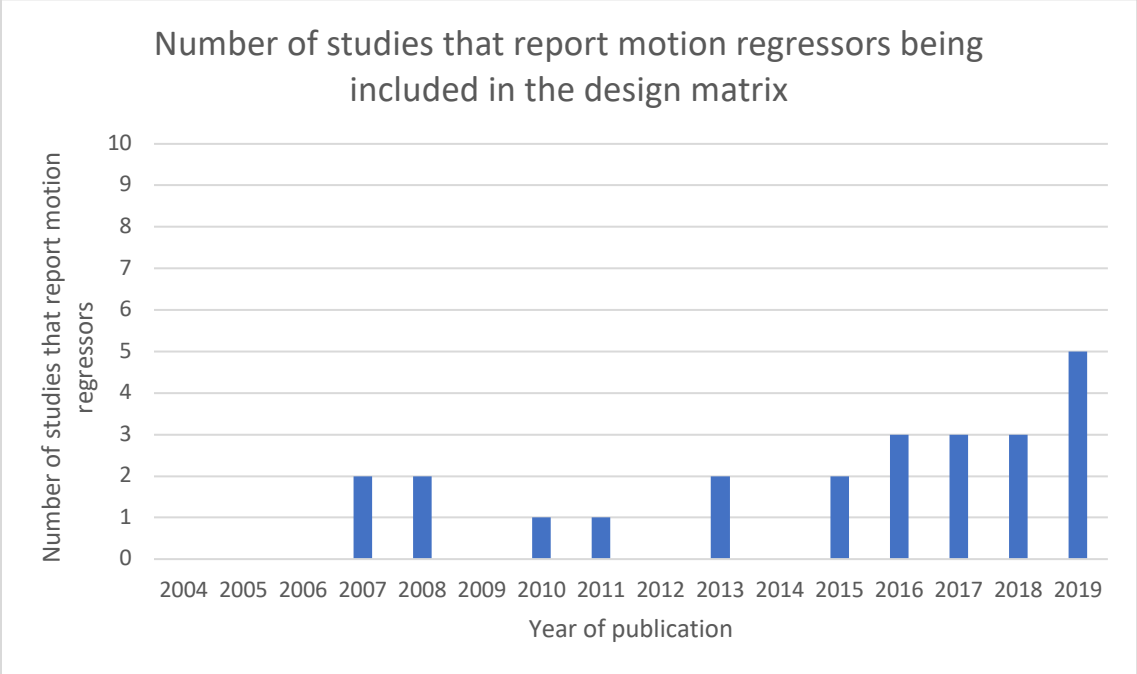

Figure 12. Reporting on the inclusion of motion regressors in the design matrix is done by a minor part of the studies.

For 28 out of 160 studies the standard space was not clearly described. There is no trend visible over the years and no statistically significant difference was found (t = -0.383, p = 0.708).

Four main standard spaces were identified in the papers: MNI (85), Talairach-Tourneaux (43), and EPI (4). In the earlier years Talairach-Tourneaux was the most popular standard space, from 2010 on MNI took over (see Figure 1313).

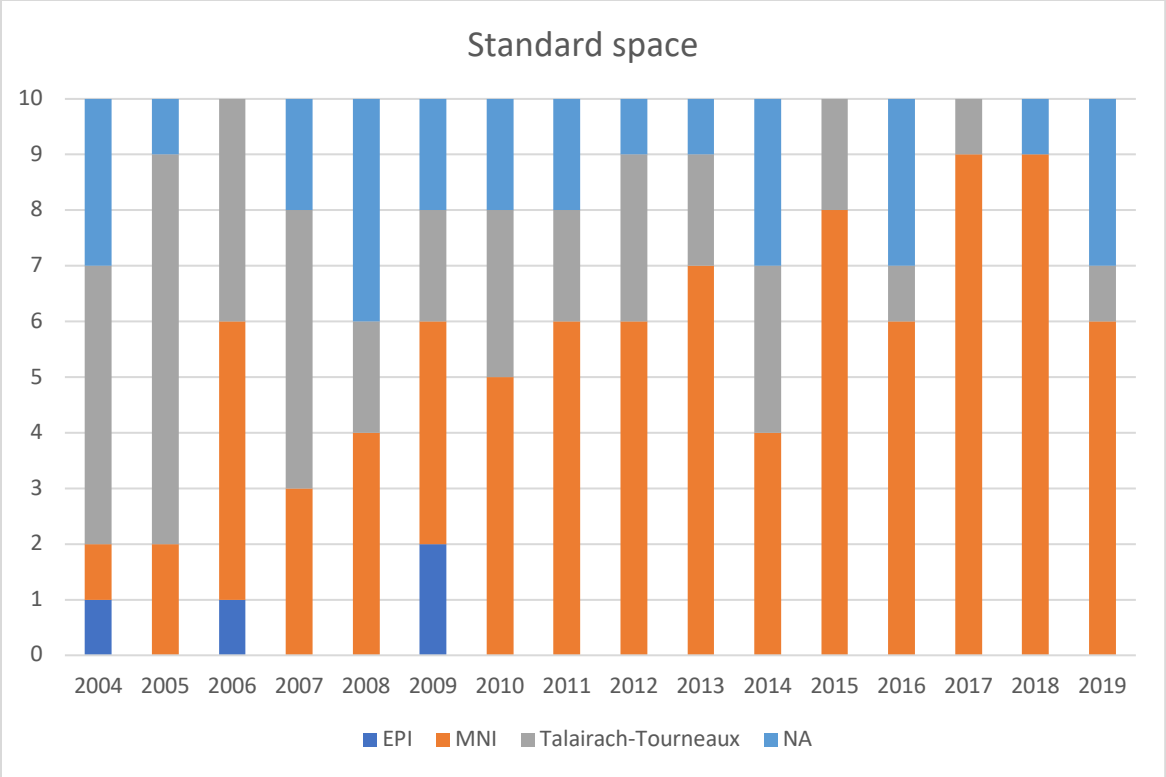

Figure 13. Evolution and distribution of the standard space that was used in the studies. The two main standard spaces are Talairach-Tourneaux and MNI. From 2010 on MNI became the most commonly used standard space.

In 90% of the studies the smoothing kernel is described. The kernel ranges between 3mm and 12mm, with 4mm, 6mm and 8mm being the most popular kernel sizes.

Data Analysis

The HRF is mentioned in 87 out of 160 studies. 73 out of those 87 studies continue to describe the type of HRF. A clear trend is visible where more studies describe the HRF in more recent years, the trend is not statistically significant ($t = 2.865$, $p = 0.125$). The lowest number of studies mentioning the HRF was in 2004 (3 out of 10 studies) and the highest number was in 2018 (8 out of 10 studies, see Figure 1414).

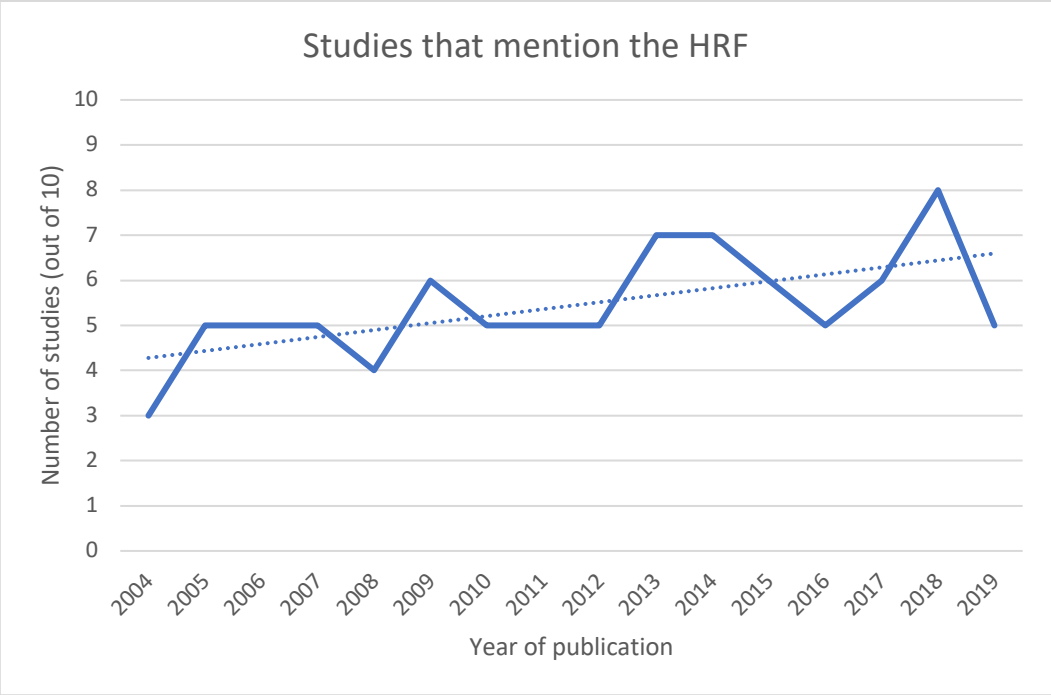

Figure 14. Evolution of the number of studies that mention the hemodynamic response function in the description of the data analysis procedure. While in 2004 only 3 out of 10 studies mentioned the HRF, in 2018 this was 8 out of 10 studies. This number varies strongly over the years, as in 2019 only 5 out of 10 studies mentioned the HRF.

In 127 out of 160 studies the interpretation of the contrast was clear. We could not see a trend over time. Only 4 out of 160 studies describe the scaling of the contrast. Furthermore, only 1 study describes the scaling of the predictors. For 120 out of 160 studies the interpretation of contrast estimates at group level is clear.

In 49 out of 160 studies the statistical model and estimation method at the first level (participant level) is not described. There is no visible trend over time, and no statistically significant difference could be detected($t = 0.276$, $p = 0.786$). In 65 out of 160 studies the statistical model and estimation method at the second level (group level) is not described (see Figure 155). The model and estimation method at the second level is less frequently reported in recent years ($t = -2.648$, $p = 0.019$).

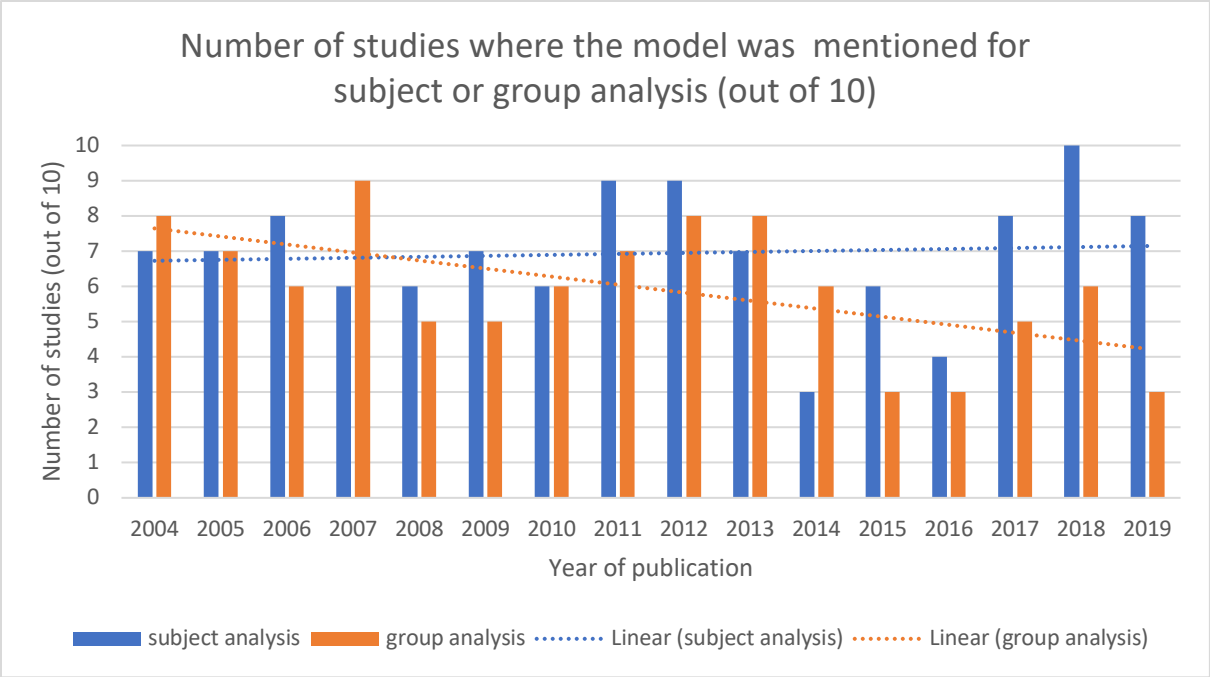

Figure 15. Number of studies (out of 10) where the model on the first level (participant level) and second level (group level) were mentioned. For the first level (participant level) we see that on average 7 studies report the model that was employed, and even though this varies a lot from year to year, there is no trend visible over the years. For the second level (group level) analysis, the method is less often reported in recent years. A lot of variability over the years is also visible here.

In 136 out of 160 studies the inference method is reported.

The employed inference methods are cluster-wise (80 studies), voxel-wise (47 studies) and mixed (7 studies). In 2 studies no group analysis was performed, these were performed in 2004. In 2004 mainly voxel-wise inference was employed, from 2005 on this changed and cluster-wise inference methods became more popular. Mixed inference methods have been reported since 2012, and voxel-wise inference methods and cluster-wise inference methods are equally employed (see Figure 16).

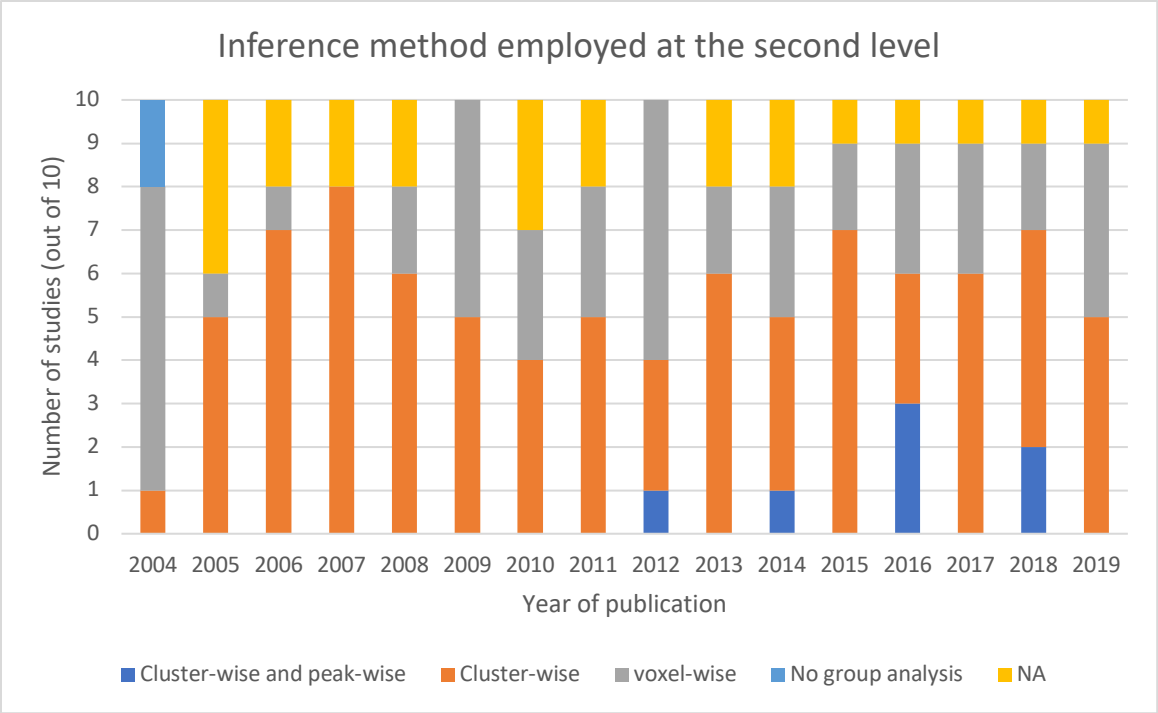

Figure 16. The distribution and evolution of inference method that was employed at the second level. In 2004 mainly voxel-wise inference was employed, from 2005 on this changed and cluster-wise inference methods became more popular.

Reporting

In 83 out of 160 studies whole brain results have been analysed and reported, 55 studies only report region of interest results (ROI) and 22 studies report both (see Figure 177). Over the years the number of studies reporting both increased, the number of studies that report whole brain results decreased from 2008 on and increased again in 2015. Subsequently, the number of studies reporting solely ROI results increased from 2008 on and has decreased since 2015.

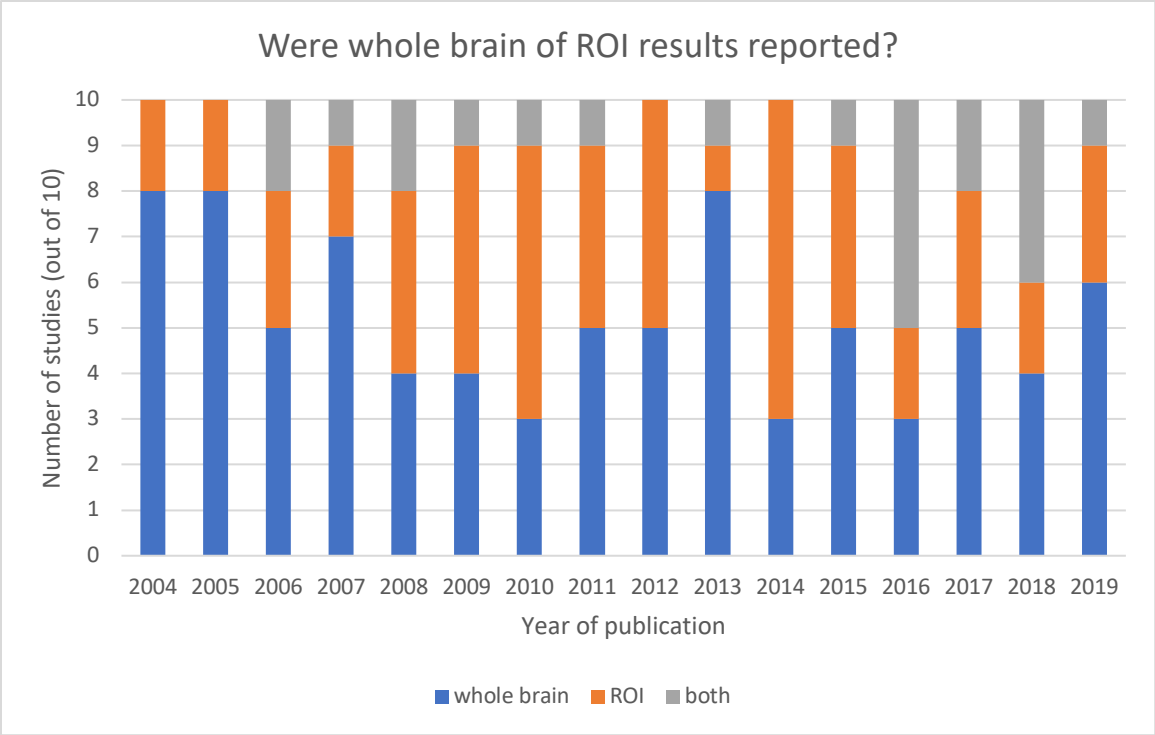

Figure 17. The number of studies that report whole brain or region of interest (ROI) results and the evolution over time. We see that reporting whole brain results became less common in more recent years and between 2009 and 2012 ROI results were the most often reported type of results.

One study out of 160 reports standardized effect sizes that are publicly available (for ROI analysis). The same study is also the only one to share statistical maps, in the form of z-values.

Ninety out of 160 studies report solely the location of local maxima in the form of xyz-coordinates, 15 studies report the combination of the location of local maxima and the statistical value of that location (see Figure 188). Finally, 55 studies report neither and therefore cannot be used for coordinate based meta-analysis methods. Most often the results are available in the form of a table, but occasionally the results are reported in a plain text.

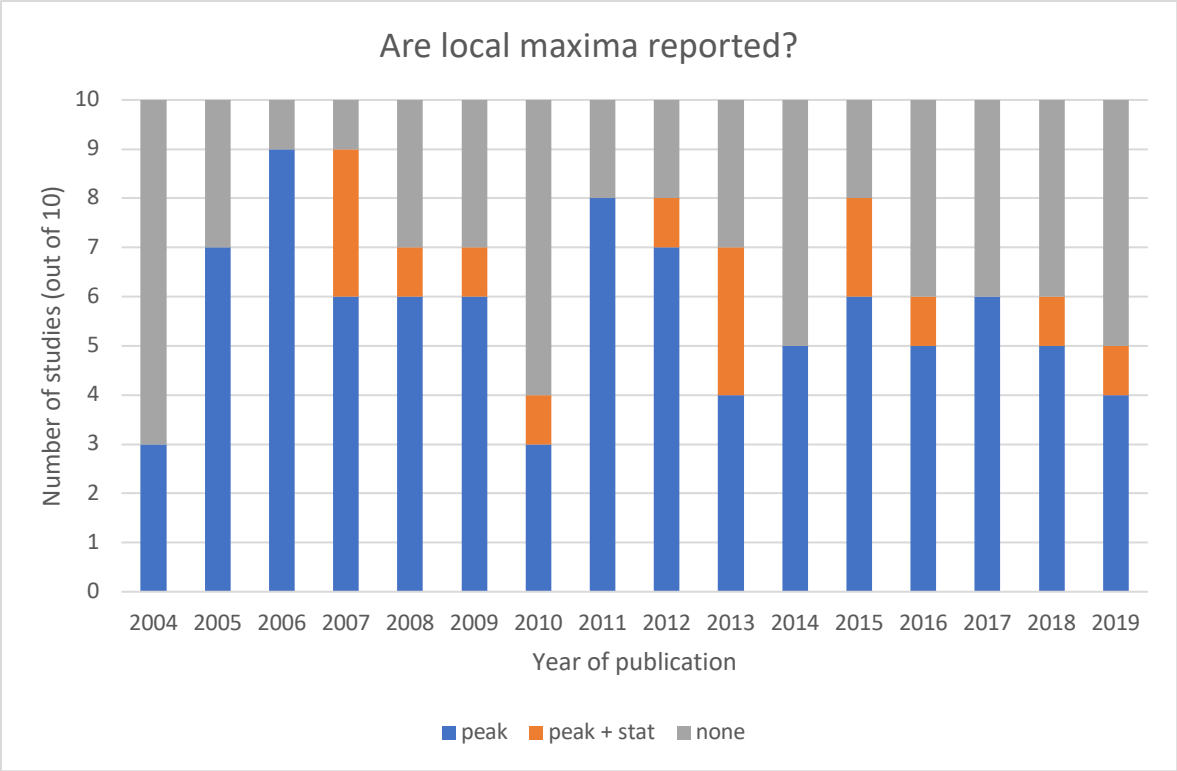

Figure 18. Evolution of how local maxima are reported of fMRI studies. When the local maxima is reported, it is often not accompanied by a test statistic. Very often no results about local maxima are reported.

Since none of the study results were published as image maps, an e-mail was sent to all authors asking whether their data could be made publicly available. For 56 studies the e-mail could not be delivered (see Figure 199). 14 authors declared that the data was not or no longer available. 11 authors indicated that the data is available if a motivated request is formulated and 1 author indicated that the data is publicly available. For 78 studies no reply was received. It was indicated that for ethical standards data had to be destroyed within a specific time frame after the study was conducted. Others indicate that there are no formal procedures to make the data available, the consent of the participants should be obtained before sharing, but there is no legal way of contacting the participants.

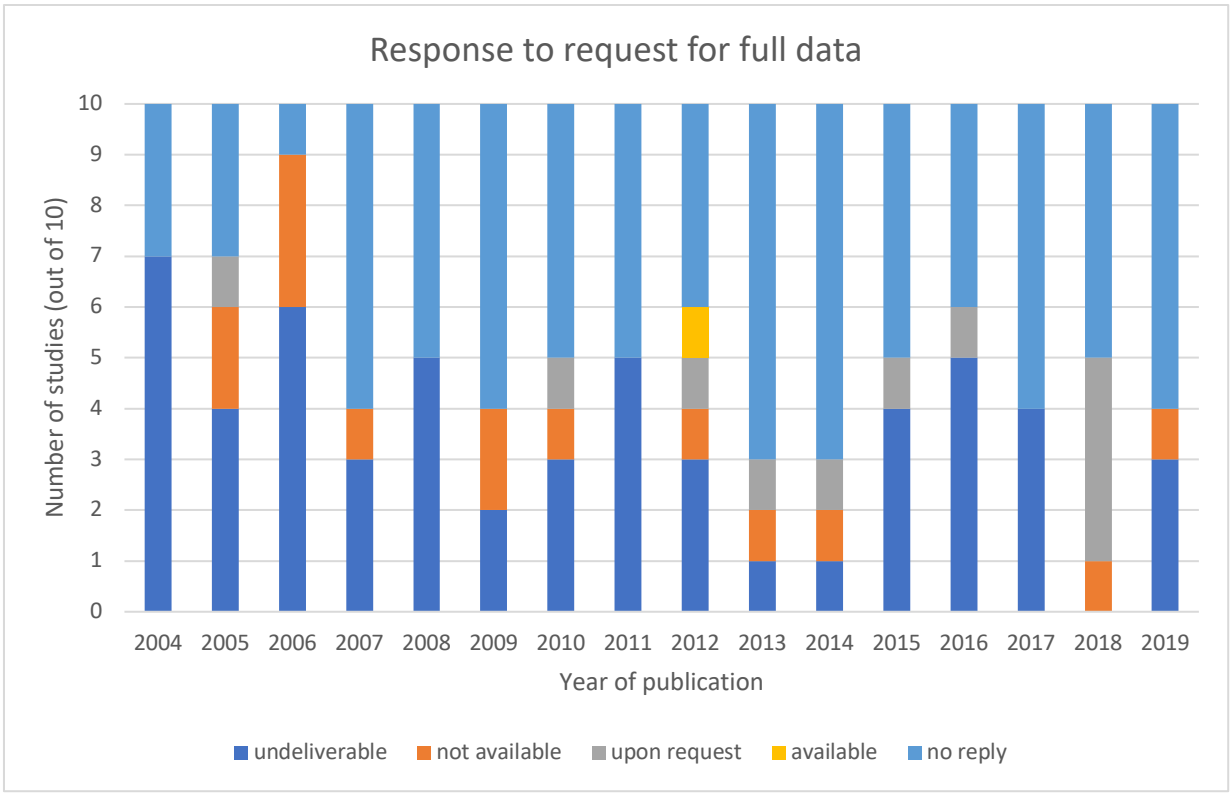

Figure 19. An e-mail was sent to all authors of the papers that were entered into this review. Most e-mails could not be delivered or received no reply. The replies that we received indicated that the results were no longer available, or were available on request with a clear research plan. Only 1 study (in 2012) indicated where the results were made available.

Discussion

In this review we describe the characteristics of fMRI studies, the way properties, choices and characteristics are reported, and their evolution over time. This is important because the way materials, methods and results are reported has an impact on how well a study can be reviewed, whether it can be replicated or not and on the eligibility of a study to be entered into a meta-analysis.

Study design

Over 85% of the studies describe the study design and there is no statistically significant trend visible over the years. When we look at the type of design we see that a block design is most frequently employed. A block design has the advantage of having higher power. A disadvantage of a block design is that it is sensitive to for instance head motion and other factors that infer with the signal. In recent years more event-related designs are reported. Event-related designs are more representative of real-life situations but they have significantly less power than block designs. It is possible that more event-related designs are used because power in studies became larger as more participants were used. Curiously, in recent years the design is less often clearly mentioned. This is different from other trends in reporting, where in more recent years a more clear reporting style is used. It is possible that this is caused by the appearance of more complex study designs that are more difficult to explain. 9% of the studies mention design optimization, which is in line with the findings of (Carp, 2012), where 5.9% of studies mentioned design optimization.

Participant characteristics

All studies report the number of participants. However, not all studies clearly describe the exclusion criteria for participants. When the exclusion criteria are not clearly mentioned it is more difficult to ascertain that participants were not excluded because they did not show desired results. This, in turn, can bias meta-analysis results. Therefore, it is vital that exclusion criteria for participants are mentioned. Fortunately, the exclusion criteria for participants, where the reasons for excluding participants from the final analysis are described, are more often clearly stated in recent years than in earlier studies, however, this trend is not statistically significant. 63% of studies describe the exclusion criteria of participants in detail, which is distinctly more than the findings of (Carp, 2012), where merely 27.9% of the studies mentioned exclusion criteria of participants.

Furthermore, the sample size clearly increases, from 15.7 participants on average in 2004 (median: 13) to 52.8 participants (median: 46) on average in 2019. This indicates that the importance of larger sample sizes for fMRI studies has become apparent and confirms earlier results where the median was 15 (Carp, 2012). Another study found a median of 34 participants per study with publication years 2010 and 2011 (Guo, et al., 2014). For comparison, we found a median sample size of 20 for these two publication years.

Next to sample size, also awareness for the importance of power to detect true activation becomes larger. Before 2010 only 3 (out of 60) studies mention statistical power in the research paper. Even if from 2010 on the majority of the papers do not mention statistical power, some awareness can be detected as studies indicate that their their study may be underpowered. Only 3 studies have performed a power analysis beforehand to determine the required sample size for the study. The high complexity and lack of tools to compute statistical power for task-based fMRI studies might be the cause of this. Since the greater part of studies that mention power mainly focus on a lack of power and that this might cause bias in the results, we can conclude that the development of intuitive and transparent tools to compute statistical power for task-based fMRI studies is imperative.

Scanner characteristics

All studies mentioned scanner field strength. However, reporting of voxel resolution and brain dimensions is not standardized. There are many different ways of reporting these characteristics, and often it is not or not unambiguously done. Most often the axial dimension is not mentioned. Likewise, it is often unclear whether the whole brain was scanned or not.

Data processing

Which data processing package was used influences the scaling of the characteristics and is vital for image-based meta-analyses. We see that 149 out of 160 studies (93%) clearly report the data processing package that was used, which is similar to the 95.4% of studies from (Carp, 2012). FSL (23 studies, 14% here and 13.9% in the study of Carp), SPM (94 studies, 59% here and 64.3% in Carp), AFNI (15 studies, 9% here and 13.9% in Carp) and BrainVoyager (17 studies, 11% and 5.7% in Carp), for 11 studies an in-house script was used or the software was unknown (7%).

Pre-processing consists of several steps. The first step is motion correction. In the scanner it is difficult to lie perfectly still. For movements to some degree motion correction can be applied to correct for these movements. The procedure that was followed for motion correction was often well described. 90% of the studies at least mention motion correction, 69% describe how the motion correction was performed. This is similar to previous results, where 94.6% of the studies mentioned motion correction (Carp, 2012).

In some studies head motion parameters are included in the GLM. 15% of studies explicitly mentioned includion motion parameters in the GLM. These studies were published from 2007 on, and th number increased after 2015.

A next step is registration to a standard space, to make the results comparable for different participants with different brains and brain sizes. Registration was also very often clearly described. 84% of the studies at least mention registration and 67% describe in detail how it was performed. This is a bit lower but still in line with previous findings, where 90.9% of the studies mentioned the normalization of images to a common template (Carp, 2012).

The third step is spatial filtering. In this step the signal is smoothed out over neighbouring voxels to remove random spikes and increase signal to noise ratio82% of studies at least mentioned spatial filtering and 78% describe the methods into detail. This is less than in a previous review where 88% of studies at least mentioned spatial smoothing of the BOLD data (Carp, 2012). Nevertheless, a large portion of studies report on spatial filtering of the data.

The fourth and final step however, temporal filtering, which is done to attenuate the signal to remove random spikes and increase signal to noise ratio was most often not clearly described. A standardized reporting format for the pre-processing steps could greatly increase the accuracy and completeness of its reporting. 48.75% of studies at least mentioned temporal filtering and 40% describe the methods into detail.

Data analysis

To model the fMRI signal in participants the hemodynamic response function (HRF) is used. Different models and applications of the HRF exist, and which one is employed evidently has an influence on the results and therefore needs to be clearly reported. A little less than half the studies give detailed information about the HRF and only a little over half of the studies even mention it. In more recent years fortunately this number increases, but this trend is not statistically significant.

Predictor and contrast scaling are complex processes, and this is reflected in how often they are described. Only 4 studies, all published between 2005 and 2007, report the contrast scaling, and only 1 study (published in 2005) reports predictor scaling.

In 69% of the studies the model for the first level (participant level) analysis is mentioned. Almost always this is a General Linear Model, sometimes a fixed effects analysis and once or twice a correlation map or a student's t analysis. There is no statistically significant difference in how well the analysis methods are described over the years for the first level. For the second level we see that in more recent years the method is less often clearly described than in the early 2000's. In 60% of the studies the model and analysis method is described.

In 15% of the studies the inference method is not explicitly mentioned. There is also a lot of variation in inference methods, ranging from very lenient uncorrected thresholding to very stringent FWE-corrected thresholding. Due to the large number of statistical tests, the low SNR and the small number of participants lenient thresholding methods are often employed to detect statistically significant results. Cluster-wise thresholding was a solution where more lenient thresholds could be employed if adjacent voxels all were statistically significant activated. Nevertheless, we see a trend in recent years to employ less cluster-wise and more voxel-level or mixed thresholding methods. The reasons for this are unclear, but it would be most valuable to investigate this further. In the earlier studies sometimes no group analysis was performed. The majority (45 of 95 reported methods) reports a random effects method, 22 report a mixed effects method and only 7 a fixed effects method. For the other 11 studies the type of method is not clearly reported. We see again the methods are more often and more clearly reported in recent years.

Reporting characteristics

Another method to more easily detect statistically significant voxels is by reducing the number of tests, for instance by analysing only a region of interest (ROI). Here you have a hypothesis beforehand about where the activation should occur. However, extrapolating the results to other studies and performing a meta-analysis is impossible if a region of interest analysis was performed. Nevertheless, we see that half of the studies have employed ROI analysis, sometimes in combination with whole brain analysis. An increase could be seen from 2008 on, but fortunately since 2015 the number of ROI analyses has decreased.

Only one study reported standardized effect sizes, and this was for a ROI analysis. Out of 160 studies, no results were available that could be entered in an image-based meta-analysis. The standard way of reporting fMRI results is still a table with the xyz-coordinates of the local maxima, often accompanied by a statistical value. Sometimes the coordinates are described in the text. These results can be entered in a coordinate-based meta-analysis. We see that 65% of the studies report the coordinates of local maxima and can therefore be entered in a CBMA. Subsequently, one third of all studies does not report any results that can be entered in a meta-analysis. This implies that for one third of the studies the results could not be used elsewhere in a quantitative manner. We did not expect it to be this many studies .

We sent an e-mail to the authors asking whether the results could be used in a meta-analysis and if the statistical maps could be shared to facilitate this. In 56/160 the e-mail could not be delivered, indicating that the main author is no longer working for the institution. This reflects the issue job insecurity poses for science, work is often not continued. 78/160 e-mails were delivered, but we received no reply. Out of the 26 remaining studies 14 authors indicated that the data was not or no longer available, for 11 studies the data was available with a motivated request after a thorough process of ethical checks. Only for 1 study was the data openly available.

There is a lot of variation in the way fMRI studies are reported. Nevertheless, more consistency in reporting styles can be observed in more recent years. Consistency in reporting style renders it more straightforward to enter studies in a meta-analysis, as it is easier for researchers to find the characteristics they are looking for and ensures that the characteristics have the same meaning across papers.

One of the main obstacles of reporting fMRI study results is the file type and size of these results. Statistical maps take up quite a lot of space. Fortunately, databases have been developed where fMRI results can be published for free. In Table 2 we list some of these places. This list is not exhaustive since for some domains, such as Alzheimer's disease, have developed separate databases.

Table 2. Databases for sharing fMRI results.

	Database	Type of data	url	RRID
1	Openneuro	Raw images or statistical maps	https://openneuro.org/	SCR_005031
2	Neurovault	Statistical maps	https://Neurovault.org/	SCR_003806
3	NITRC	Statistical maps	https://www.nitrc.org/	SCR_003430
4	OSF	Statistical maps	https://osf.io/	SCR_003238
5	Harvard dataverse	Statistical maps	https://dataverse.org/	SCR_001997
6	Brainspell	Coordinates of local maxima	https://brainspell.org/	SCR_001639
7	BrainMap	Coordinates of local maxima and sample size	https://brainmap.org/	SCR_003069
8	Paper	Coordinates of local maxima and sample size		

In this paper we focus on reporting practices for task fMRI studies. With fMRI it is also possible to perform resting state studies, where no task or paradigm is presented to the participant and brain activity is measured while the participant is "resting". Resting state methods are typically more complicated than task based fMRI studies and therefore require even more exhaustive reporting practices. While we focus on task fMRI in this paper, it would be valuable to investigate reporting practices for methods and results of resting state fMRI studies.

The main limitation of this study is the number of published studies that are included in the review. Even though 160 studies were included, this still refers to 10 studies for every year of publication. Nevertheless, we believe that the sample size is sufficiently large to get a general idea of reporting styles for methods and results of task fMRI studies.

Furthermore, while most criteria were evaluated in a quantitative manner, some were qualitatively evaluated because of the large variation in reporting methods in the individual studies. This leaves room for interpretation. However, we believe that including these qualitative measures was valuable and we rather include them in a manner that leaves room for interpretation, than not at all.

Finally, the parameters and characteristics that are included in this review are not exhaustive. Reporting practices can be studied in more detail. In this review we made a selection of parameters that we would study, based on how relevant they are for comprehending and replicating the study, or including it in a meta-analysis.

Some parameters were not included, for example how many studies collect field maps or collect volumes with different phase encoding directions. Which option is chosen defines whether and how can be corrected for geometric distortions. Nevertheless, we note that such specific information is

rarely written down in a research paper. A solution to this is BIDS. If in the future researchers export their analysis settings through BIDS the comprehension, reproduction and inclusion in meta-analyses of task fMRI studies will be greatly facilitated.

Conclusion

In this paper we have reviewed reporting styles and methods of fMRI studies. We found that characteristics with regards to the method of a study are more often mentioned in recent years compared to earlier years. With regards to the reporting of results we found that the location of local maxima is most often shared in the form of coordinates, sometimes accompanied by a test statistic. Statistical maps are very rarely shared, and only seldomly available upon request. What type of information about study methods and results and the way it is reported has a large impact on the eligibility of a study to be entered into a meta-analysis. In a follow-up manuscript, we explore the impact of reporting styles on the eligibility for meta-analysis techniques where we dive deeper into meta-analysis methods and the input they require.

References

Bowring, A., Maumet, C., & Nichols, T. (2019). Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping*, Vol 40 (11), p 3362-3384.

Carp. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage, 63*, 289-300.

Gorgolewski KJ, V. G.-B. (2015). NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the brain. *Front. Neuroinform.* doi:doi: 10.3389/fninf.2015.00008

Guo, Q., Parlar, M., Truong, W., Hall, G., Thabane, L., McKinnon, M., . . . Pullenayegum, E. (2014). The Reporting of Observational Clinical Functional Magnetic Resonance Imaging Studies: A Systematic Review. *PLOS One*, e94412.

Kao, M.-H., Temkit, M., & Wong, W. K. (2014). Recent developments in optimal experimental designs for functional magnetic resonance imaging. *World Journal of Radiology*, 6(7): 437–445.

Larrazabal, A. J. (2020). *Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis.* Proceedings of the National Academy of Sciences of the United States of America, 117(23), 12592–12594.

Logothetis. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*, 869-878.

Markiewicz, G. F. (2021). The OpenNeuro resource for sharing of neuroscience data. *eLife*. doi:10.7554/eLife.71774

Mehta, R., & Parasuraman, R. (2013, 12). Neuroergonomics: A Review of Applications to Physical and Cognitive Work. *Frontiers in human neuroscience, 7*, 889. doi:10.3389/fnhum.2013.00889

Nichols. (2012, July 31). *SPM plot units*. Retrieved from Warwick blogs: https://blogs.warwick.ac.uk/nichols/entry/spm_plot_units/

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., . . . Yeo, T. B. (2016). *Best Practices in Data Analysis and Sharing in Neuroimaging using fMRI.* http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf: Human Brain Mapping. Retrieved from http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf

Poldrack, Baker, Durnez, Gorgolewski, Matthews, Munafò, . . . Yarkoni. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*, 115-126.

Poldrack, Fletcher, Henson, Worsley, Brett, & Nichols. (2008). Guidelines for reporting an fMRI study. *Neuroimage*, 409-414.

Conflict of Interest

The authors do not wish to state any conflict of interest. The data can be found in https://osf.io/5swve/ . The study has not been preregistered and the measures and analyses performed as part of this study have been reported and can be found in the data file.

Appendix A. Table with overview of all variables that were registered.

Identification	PubMedID	ID of the publication on PubMed
	pub_year	Year of publication
	Title	Title of the pulication
	include?	Do we include this paper, and if not, relevant exclusion criterium
Design	Design	Is the design paradigm described?
	Type of design	Block/event related
	Optimalisation	Was the design optimized? For instance, were the stimuli jittered? (0 = no, 1 = yes)
	Multiple experiments?	Was more than one experiment included in the paper? (0 = no, 1 = yes)
	Was this the main experiment?	Was the fMRI experiment and univariate approach the main focus of the paper? (0 = no, 1 = yes)
Study characteristics	N	Number of participants
	Gender v	Number of female participants
	Gender m	Number of male participants
	Ratio gender	If equal or close to 1 = balance, >1 more women, <1 more men
	Exclusion crit	Are the exclusion criteria for participants clearly described in the paper? (0 = no, 1 = yes)
Scanning	T	Field strength of the scanner in Tesla (list: 1.5, 3, 4, 7)
	Whole brain meas	Was the whole brain scanned? (0 = no, 1 = yes)
	Vox res	Voxel resolution (number x number x number in mm)
	Image dim	Dimensions of the scanned image (number x number x number in voxels)

Processing the data	Software	Software + version number
	Motion correction	Mentioned? 0 = no, 1 = mentioned, 2 = clearly described and can be reproduced
	Registration	Mentioned? 0 = no, 1 = mentioned, 2 = clearly described and can be reproduced
	Spatial Filtering	Mentioned? 0 = no, 1 = mentioned, 2 = clearly described and can be reproduced
	Temporal Filtering	Mentioned? 0 = no, 1 = mentioned, 2 = clearly described and can be reproduced
	Preprocessing	Score for how well the preprocessing has been described, sum of previous three
	Coordinate space	Which coordinate space was used?
	Smoothing (FWHM)	FWHM of the smoothing kernel (number)
Data analysis	HRF model	Is the HRF model mentioned? (0 = no, 1 = yes)
	HRF model	If yes, which model was used for the HRF (text)
	Contrast	Was the contrast clearly described? (0 = no, 1 = yes)
	Contrast scale	Was the scaling of the contrast reported? (0 = no, 1 = yes)
	Predictor scaling	Was the scaling of the predictors reported? (0 = no, 1 = yes)
	Contrast group level	Is the interpretation of contrast estimates at group level clear? (0 = no, 1 = yes)
	Participant analysis	What is the statistical model and estimation method used for the first level?
	Group analysis	What is the statistical model and estimation method used for the group analysis?
	Inference method	Cluster, peak, … NA if not mentioned

	Inf spec	Specific details about the inference method
	Cluster forming threshold	In case of topological inference, was cluster forming threshold specified? (give values)
	Thresholding method	Which method was used, what was the threshold? (NA if not mentioned)
Reporting	Whole brain	Were results analyzed and reported on whole brain level? (0 = no, 1 = yes)
	Standardized effect size	Are maps with standardized effect sizes available? (0 = no, 1 = yes)
	Stat maps	Are statistical maps available? (0 = no, 1 = yes)
	Share maps	How were the maps shared? (after e-mail, in which database, …)
	test statistic	Is a z/t/F-map shared? (0 = no, 1 = yes)
	contrast/parameter estimates	(0 = no, 1 = yes)
	standard errors	(0 = no, 1 = yes)
	Peak stat	Were local maxima with statistical value shared? (0 = no, 1 = yes)
	peak	Were only the local maxima shared? (0 = no, 1 = yes)
	Data after e-mail	Was the data available after an e-mail was sent to the authors? (0 = no, 1 = for collaboration, 2 = could be made publicly available, -1 cant reach author)

Appendix B. Frequency of use of analysis software

Software	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
AFNI	3	0	0	1	1	1	1	1	2	0	2	1	0	0	1
BrainVoyager	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
BrainVoyager 2000	1	2	1	1	0	1	1	0	0	0	0	0	0	0	0
BrainVoyager 4.8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
BrainVoyager QX	0	0	1	1	2	1	0	0	1	1	0	1	0	0	0
Own software	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
FSL	0	0	0	1	3	3	0	2	1	3	1	2	2	1	4
IDL	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
MATLAB	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
MEDx3.2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

NA	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0
SPM	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SPM12	0	0	0	1	0	0	0	0	0	0	0	0	0	2	2
SPM2	0	0	2	0	3	4	4	1	0	1	2	0	0	1	0
SPM5	0	0	0	0	0	0	1	4	2	0	2	2	3	1	0
SPM8	0	0	0	0	0	0	1	1	3	4	3	4	5	5	3
SPM99	6	4	6	3	1	0	1	0	1	0	0	0	0	0	0
XBAM	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0