Assessing Indices of Transparency and Reproducibility in Animal Models of Opioid Addiction

315

University of Sydney

Empirical Thesis submitted in partial fulfilment of the requirements for

Honours in Psychology at the University of Sydney, 2023

Word Count: 11941

Table of Contents

List of Figures

List of Tables

Abstract

Psychology's reproducibility crisis has led to a reckoning of research practices in many fields. Moreover, several preclinical fields have come under scrutiny due to poor rates of treatment translation from animals to humans. This is true of the preclinical addiction field (Venniro et al., 2020). Ensuing investigation revealed that many of the same research design aspects that undermine reproducibility also threaten translation potential (Fergusson et al., 2019). We examined indices of transparency and reproducibility in animal models of opioid addiction from 2019 to 2023. In doing so, we aimed to understand whether efforts to improve reproducibility are relevant to this field. We measured the prevalence of transparency measures such as preregistration, registered reports, open data, and open code as well as compliance to the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines. We also measured reported rates of bias minimisation practices, sample size calculations and multiple corrections adjustments. Lastly, we estimated the accuracy of test statistic reporting. Appraising 247 articles revealed poor uptake of transparency measures, the ARRIVE guidelines, bias minimisation practices and sample size calculations. Adjustments for multiple comparisons was alone in being implemented in most articles (76.5%). Lastly, half of articles contained non-decision errors and 11% contained decision errors. We discuss the implications of these results and potential explanations as well as solutions for their improvement. Our study is the first of its kind in this field and demonstrates that attempts to improve reproducibility and, in turn, translation, are needed in the animal models of opioid addiction field.

Keywords: animal models of opioid addiction, reproducibility, translation, transparency, bias minimisation, accuracy, reporting standards, ARRIVE

Assessing Indices of Transparency and Reproducibility in Animal Models of Opioid Addiction

The development of treatments for human psychopathology is one goal of psychological research. Research that is accurate and unbiased is valuable to this pursuit as it is more likely to lead to reliable and effective interventions (Landis et al., 2012; Schmidt-Pogoda et al., 2020). Accuracy in research, however, can be difficult to ensure.

Verifying findings through the replication of studies is one measure of accuracy (Nosek et al., 2012). Because science is built on verifiability, validating results in this way should be common practice and an integral part of any research culture (Munafò et al., 2017). Good reproducibility in a field indicates robust findings, a worthy investment of funding and resources, and low rates of unnecessary risk to humans and non-human animals (hereafter, animals).

Large scale replication attempts in psychology aimed to assess the field's reproducibility. The results were alarming: only 39% of replications were considered successful (Open Science Collaboration, 2015). Further, the studies that replicated produced effect sizes on average half the magnitude originally reported (Open Science Collaboration, 2015). This *reproducibility crisis* triggered widespread discussion about how psychological research is performed (Pashler & Wagenmakers, 2012).

Researchers in other disciplines have called for similar replication efforts to address potential shortcomings in their own fields: appeals beginning in gambling addiction research have spread to include addiction research more broadly (Heirene, 2021).

In preclinical animal research, the goal is translation: the successful application of results from animals to humans is the goal. Successful translation can lead to effective treatments for human addictions. A failed translation can indicate many things, for example, the validity of the animal model, the efficacy of the treatment or subpar preclinical research methodology. It takes approximately US$330 000 and the time of - often ill - people to see if an intervention will translate (Perrin, 2014). Given this investment, it is imperative that poor methodology in the preclinical stages can be ruled out as a cause in the case of a failed translation (Kimmelman & Anderson, 2012). Unfortunately, this is often not possible (Perrin, 2014; Schulz et al., 2016).

Disappointing levels of translation to clinical trials – including in addiction research – have led some to declare a *translation research crisis* (Perrin, 2014; Venniro et al., 2020). Importantly, there is overlap between the contributors to the translation research crisis and the replication crisis (Fergusson et al., 2019). Randomisation, masking, and data exclusion, as well as researcher misconduct and systematic influences have been discussed as problem areas in both crises (Landis et al., 2012; Munafò et al., 2017). Thus, successful translation is supported by many of the same rigorous research practices that underpin successful replication (Schulz et al., 2016).

The ongoing opioid epidemic in North America and Australia, among other countries, has led to an enormous loss of life (Australian Institute of Health & Welfare, 2018; National Institutes of Health, 2023b). A better understanding of the mechanisms that underly opioid addiction, treatments for opioid addiction, and non-addictive analgesia alternatives is needed (Epstein et al., 2018; National Institutes of Health, 2023a). Animal models of opioid addiction (AMOA) research that is transparent and reproducible is the solid foundation from which translatable treatments may be developed. However, to date, there has been no investigation into the reported prevalence of research practices that support these processes in the AMOA literature.

It should be noted that there is no consensus on the existence of either the replication crisis or the translation research crisis. The nomenclature, however, may be beside the point. What is of importance is that there is significant room for improvement in terms of translation and reproducibility in multiple preclinical fields (Landis et al., 2012; Macleod et al., 2015). Understanding the extent to which proposed solutions to these crises are relevant to the field of AMOA is the motivation of the current study.

Causes of the replication crisis

At the systematic level, publication bias has been discussed as one cause of poor rates of reproducibility and translation (Landis et al., 2012; Open Science Collaboration, 2015). Publication bias describes the tendency for journals to publish papers with significant findings over non-significant findings, and to favour ‘tidy’, linear studies that culminate in novel discoveries (Giner-Sorolla, 2012). This bias is responsible for disproportionate levels of false positives and inflated effect sizes in the literature (Simmons et al., 2011). In preclinical stroke research, effect sizes were estimated to be inflated by 30% due to publication bias (Sena et al., 2010). Further, publication bias leads to a literature that is not representative of the entirety of the research being done (Moher et al., 2016). In animal research, it is estimated that only 60-67% of research carried out is published (van der Naald et al., 2020). In preclinical research, these factors preclude clinicians from making informed decisions about which treatments to progress to clinical trials (Kimmelman & Anderson, 2012; Moher et al., 2016).

Importantly, publication bias incentivises researchers to find statistically significant results (Munafò et al., 2017). This may lead researchers to – wittingly or unwittingly – engage in questionable research practices (QRPs) to achieve significant results (John et al., 2012; Simmons et al., 2011). Indeed, undisclosed QRPs were found to be surprisingly common in psychology researchers from a wide range of disciplines (John et al., 2012). Some common QRPs are described in Table 1.

QRPs undermine the main goal of scientific research: to accurately describe true effects. Furthermore, because the existence of a publication bias means scientific journals are unlikely to publish replications and non-statistical findings, research that challenges published positive findings has very little chance of publication (Antonakis, 2017). This means that false positives are hard to correct (Simmons et al., 2011). Furthermore, while random bias can be removed by aggregating data, systematic bias cannot (Scheel et al., 2021). This means that meta-analyses are unable to correct for a biased literature of potentially inflated effect sizes (Scheel et al., 2021). This indicates that solutions to these problems must be largely preventative.

Reproducibility

The term ‘reproducibility’ has been described as ‘overloaded’ as there are distinct, though related, types of reproducibility (Stodden et al., 2013). What follows is a brief discussion of some of the relevant kinds of reproducibility.

Firstly, *results reproducibility* describes lab collecting and analysing new data following the methodology of an original paper (Goodman et al., 2016). A successful results replication provides support for the reliability of an effect (Nosek et al., 2012). An unsuccessful attempt, on the other hand, can indicate many issues: an unfaithful replication, the absence of an effect, or unsound methodology in the original paper (Open Science Collaboration, 2015)

Replication attempts in psychology spurred similar attempts in other fields. The Replication Project: Cancer Biology attempted replications in preclinical cancer biology. They considered 46% of original effects to have successfully replicated (Errington, Mathur, et al., 2021). This suggests poor results reproducibility is not a concern for soft sciences alone. Indeed, there have been calls for similar efforts in addiction due to concern over the field’s reproducibility (Heirene, 2021).

Naturally, large-scale replication efforts may be unnecessary in fields that already have high rates of replication. Despite low rates of replication in addiction research more broadly, AMOA may be such a field (see Table 2 for relevant estimates) (Adewumi et al., 2021). This idea finds support in the fact that replication is valued in preclinical research, as evinced by the use of biological and technical replicates (Lazic et al., 2018). We will assess the rate of results replications in the current study to answer this gap in knowledge.

The second type of reproducibility is referred to as *computational reproducibility*. This involves rerunning the analysis code on the original data and therefore is predicated on access to these materials. Computational reproducibility is an efficient way of verifying results and helps to rule out errors in statistical analysis as a reason for failed results replication (Eubank, 2016). Computational reproductions have revealed inaccuracies in statistical analyses, from inconsequential errors to decision errors and inaccurate effect size estimations (Eubank, 2016; Hardwicke et al., 2018). Failed computational reproducibility weakens the credibility of a paper's findings by revealing that they cannot be substantiated by the original data and analyses (Hardwicke et al., 2018). Consequently, it is considered the 'minimum level of credibility' a field would hope to have (Eubank, 2016; Hardwicke et al., 2018).

Thirdly, *methods reproducibility* asks if there is enough methodological information provided to attempt a replication (Goodman et al., 2016). This type of reproducibility relies on detailed reporting practices, which is also crucial for the reader to be able to adequately judge the validity and reliability of a paper's results (Percie du Sert, Hurst, et al., 2020). However, a lack of thorough reporting remains a roadblock: the initial attempt in the Reproducibility Project: Cancer Biology was unable to replicate any experiments due to incomplete methods reporting (Errington, Denis, et al., 2021).

Solutions have been proposed for improving these different types of reproducibility. Typically, these solutions centre on increasing transparency in research and ameliorating reporting standards. To this end, the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines were developed (Percie du Sert, Hurst, et al., 2020). It is hoped that improving reporting will encourage rigorous and transparent research, thus leading to more robust findings, better reproducibility and ameliorated translation rates (Fergusson et al., 2019; Percie du Sert, Hurst, et al., 2020)

Currently, metascience researchers are engaged in examining transparency and reporting practices in different fields to understand their prevalence. By doing this in AMOA for the first time, we hope to understand to what extent efforts to improve reproducibility – and, in turn, translation – are relevant to this field.

Table 1

Questionable Research Practices Contributing to the Reproducibility Crisis

Name	Definition	Implications	Consequences	Associated target variables [a]
HARKing	The researcher adjusts their *a priori* hypotheses after seeing results to more accurately 'predict' the study's outcome (Kerr, 1998)	Exploratory research ("we are not sure what is happening here so we will test for a few things to try and find out") is misrepresented as confirmatory research ("we think there is X effect here, we will test for it")	The evidential weight for an effect is overestimated (Kerr, 1998)	Preregistration Registered reports
p-hacking	The undisclosed omission, transformation, or combination of variables until statistical significance is reached (Simonsohn, 2014)	By retesting the hypothesis multiple times, *p*-hacking violates the assumptions of null hypothesis significance testing (Simmons et al., 2011)	Validity of results undermined (Simmons et al., 2011) Likelihood of finding a false positive increases (Simmons et al., 2011)	Preregistration Registered reports Open data Open code
Outcome switching	Swapping the variables of interest in a study after seeing the results, often to reach significance (Vassar, Roberts, et al., 2020)	Misrepresents efficacy of a treatment at preclinical or clinical stages (Vassar, Roberts, et al., 2020)	Mislead future clinical research (Vassar, Roberts, et al., 2020) Precludes clinicians from making fully-informed decisions about	Preregistration Registered reports

Name	Definition	Implications	Consequences	Associated target variables [a]
	May occur consciously or unconsciously (Munafò et al., 2017)		preclinical treatment efficacy (Kimmelman & Anderson, 2012; Moher et al., 2016) Increased risk of finding a false positive (Simmons et al., 2011) Not considering totality of results (Munafò et al., 2017)	Open data Open code
Selective reporting	Omitting variables after observing the results, often in order to achieve significant results (Vassar, Roberts, et al., 2020) Can also be referred to as underreporting when analyses, experiments, or subjects are omitted (van der Naald et al., 2020)	Misrepresents efficacy of a treatment at preclinical or clinical stages (Vassar, Roberts, et al., 2020) Data aggregation efforts (meta-analyses, systematic reviews) cannot include all data generated for a certain outcome (Vassar, Roberts, et al., 2020)	Precludes clinicians from making fully-informed decisions about preclinical treatment efficacy (Kimmelman & Anderson, 2012; Moher et al., 2016) Mislead future clinical research (Vassar, Roberts, et al., 2020)	Preregistration Registered reports Open data Open code

[a] These variables are being examined in this study. They can help to detect the presence of a questionable research practice or mitigate its impact.

Transparency

Preregistration

Preregistration of empirical research is widely considered to be a crucial part of any solution to the reproducibility and translation crises (Gorman, 2019; Munafò et al., 2017; Nosek et al., 2019; Pennington, 2023; Schäfer & Schwarz, 2019; Scheel et al., 2021; van der Naald et al., 2020). Preregistering a study involves posting the hypotheses, research design and planned statistical analyses on an online repository. Preregistration has several benefits.

Firstly, preregistration helps to detect changes in the research plan. This means HARKing becomes obvious as readers can compare preregistered and published hypotheses (Bergkvist, 2020). Similarly, it is easier to demarcate planned from *post hoc* analyses, meaning it is clear which analyses are confirmatory and which are exploratory. As with HARKing, this distinction has implications for the strength and interpretation of the results (Simmons et al., 2011).

Clearly, departure from preregistration does not immediately indicate the presence of QRPs, but it can: a comparison of addiction randomised control trials (RCTs) to their preregistrations revealed 29% contained instances of outcome switching or selective reporting, and only 2% of these discrepancies were noted in the final paper (Vassar, Roberts, et al., 2020). Unsurprisingly, the researchers believed the discrepancies to be motivated by a desire to achieve statistical significance or to 'obscure' non-significance.

Preregistration can also mitigate publication bias by making it easier to detect instances where experiments have been left out of the published report due to non-significance (van der Naald et al., 2020). This makes all planned research 'discoverable' and may be informative for clinicians deciding which treatments to pursue (Moher et al., 2016; Nosek et al., 2019).

Importantly, subpar research practices may occur deliberately or because of unconscious biases (Munafò et al., 2017). In this way, preregistration also assists well-intentioned researchers to avoid the effects of biases.

Table 2

Prevalence of Transparency and Reproducibility Practices from Previous Work in Other Fields

		Study					
Study Characteristic		Adewumi et al. (2021)	Norris et al. (2021)	Hamilton et al. (2023)	Hardwicke et al. (2020)	Makel et al. (2012)	Pui Yu Lee et al. (2022)
Field		Addiction	Addiction (Smoking)	Health and medicine meta-research	Social sciences	Psychology	Psychology
Human or animal or both		Both[a]	Human	Both	Human, non-animal/human	Unclear	Unclear
Publishing years of papers reviewed		2014-2018	2018-2019	1781-2022 (interquartile range 2012-2018)	2014-2017	1900-2012	2010-2021
Total number of papers		244	100	105 meta-research (2 121 580 articles)	156	500	84 834
Replications		.4%	0%			1.1%	0.2%
Preregistration (%)	States preregistered	2.9%	73%		0%		

		Study					
Study Characteristic		Adewumi et al. (2021)	Norris et al. (2021)	Hamilton et al. (2023)	Hardwicke et al. (2020)	Makel et al. (2012)	Pui Yu Lee et al. (2022)
	States not preregistered	0%			0%		
	No statement	97.1%			100%		
Data availability (%)	States available (and accessible)	11.5% (8.2%)	7%	8% (2%) [b]	7%		
	States not available	2.05%			0.6%		
	No statement	87%	93%	92% [b]	92.3%		
Code availability (%)	States available	0.8%	1%	0.5% [c]	1.3%		
	No statement	99.2%	99%	99.5%	98.7%		

Note. The first four rows describe characteristics of the previous works. The following rows present their results. RCT: Randomised Control Trial.

[a] The distribution of studies was 225 human and 19 animal

[b] This figure from papers published between 2016 and 2021

[c] This figure from papers published between 2016 and 2022

Preregistration helps to detect poor research practices, while also encouraging good practices. Preregistration platforms prompt consideration of crucial research design aspects which may improve the rate at which these practices are implemented and reported (Nosek et al., 2019).

A combination of these aspects may account for the lower rates of positive results and smaller effect sizes in preregistered studies (Open Science Collaboration, 2015; Schäfer & Schwarz, 2019).

Similar results were found when comparing registered reports to non-registered reports (Scheel et al., 2021). Registered reports involve a journal accepting a paper based on the motivation and methodology before the data has been collected. In this way, registered reports have the additional benefit of combatting publication bias (Ellis, 2022).

Because of the advantages preregistration offers, high prevalence rates in a literature can be seen as an indicator of a field's robustness against threats to reproducibility and translation potential (Nosek et al., 2018).

As the first study to investigate levels of preregistration in AMOA and with mixed findings in other areas (see Table 2), we have no clear expectations. The compulsory preregistration of studies in clinical addiction research may serve to facilitate the uptake of the practice in preclinical addiction research (Munafò, 2015; Norris et al., 2021).

Open data

Other practices to improve transparency of research practices include sharing of data and code.

Open data refers to the practice of making an experiment's raw data available. A study's data is the evidence that substantiates its conclusions; as such, being unable to corroborate evidence reduces the credibility of a study's claims (Hardwicke & Ioannidis, 2018).

Open data is the first step towards computational replications which allows for verifiability of results (Hardwicke et al., 2018). Data sharing means researchers can ask related questions of the same data, thus avoiding unnecessary research duplication, saving resources, and improving efficiency of research (Hardwicke & Ioannidis, 2018; Ting et al., 2015). Further, open data encourages discourse between researchers and accelerates synthesis of evidence, ultimately benefitting a cumulative science (Eubank, 2016; Pennington, 2023).

Lastly, open data may make selective reporting more detectable (van der Naald et al., 2020). This practice is of particular concern in animal research where a considerable amount of research is not reported and therefore wasted (Moher et al., 2016)

Despite the benefits, data sharing in addiction research appears rare (Table 2). A review of addiction RCTs found zero instances of data sharing (Vassar, Jellison, et al., 2020). On the other hand, rates of data sharing in animal addiction research may be considerably higher given it is not constrained by privacy laws (Kimmelman & Anderson, 2012). Furthermore, animal addiction research utilises data repositories, such as genome or protein databases, indicating an existing familiarity with the benefits of data sharing (Munafò, 2015). These considerations may prove beneficial for open data adoption in AMOA.

Open code

Open code describes the practice of sharing the analysis script used to analyse an experiment's data. Access to a study's data and code is essential for computational reproductions. This can partly assess the reliability and accuracy of a study's results (Eubank, 2016; Hardwicke et al., 2018). Indeed, open code was seen to improve computational reproductions by 40% (Laurinavichyute et al., 2022).

While open code allows for transparency about the analytical pipeline and therefore facilitates scrutiny like preregistration and open data, open code additionally makes computational reproduction efficient (Eubank, 2016; Hardwicke et al., 2018)(. Recreating results without the code has been described as building flat-pack furniture without instructions – that is, time-consuming and difficult (Hardwicke et al., 2018). Improved efficiency of computational reproducibility is likely to increase the frequency that study results are verified (Eubank, 2016).

Furthermore, rerunning analyses can highlight issues with data formatting or labelling, thus improving the future functionality of the dataset (Hardwicke et al., 2018).

However, open code tends to be less common than open data, less frequently stipulated as a journal requirement and examined in metascience research less often (Table 2) (Hardwicke et al., 2020; Stodden et al., 2013). Indeed, while statements regarding preregistration and open data are included in ARRIVE's 'Recommended Set', open code is not mentioned (Percie du Sert, Hurst, et al., 2020). With this in mind, we do not expect code sharing to be a popular practice.

Reporting Standards

Masking

Masking – also known as blinding – involves the researcher being unaware of the group allocation of an animal during an experiment. Ideally, masking is implemented at various stages throughout an experiment (Percie du Sert, Hurst, et al., 2020). Masking is essential to avoid researcher bias influencing a study's outcome and is therefore crucial in hypothesis-testing research (Karp et al., 2022). By protecting against bias, masking improves the validity of a paper's results and the predictive value in future clinical trials (Karp et al., 2022; Watzlawick et al., 2019).

Research reveals that the absence of masking inflates effect sizes and increases the risk of false positives (Bebarta et al., 2003; Watzlawick et al., 2019). One systematic review found that a lack of blinding in clinical randomised control trials (RCTs) increased the odds ratio of the treatment efficacy by 36% compared to blinded results (Hróbjartsson et al., 2012). In this way, biased results in preclinical research have the potential to misdirect future research and are less likely to translate into effective treatments for humans (Schmidt-Pogoda et al., 2020; Watzlawick et al., 2019).

We will examine the prevalence of masking in AMOA to assess if this field may be at risk of similar issues. Given the novelty of this research, we have no clear expectations. Estimates from preclinical fields indicate generally low rates (Table 3), with a large-scale survey of preclinical biomedical literature revealing an estimate of 12.3% (Menke et al., 2020). A somewhat larger estimate of 43% was found in a survey of analgesia, anaesthesia, and animal welfare (Leung et al., 2018). This estimate may be instructive, given there is overlap in the search for opioid alternatives for pain relief.

Moreover, the evidence of improving rates of reported masking may place AMOA in line with the larger existing estimates (Kousholt et al., 2022; Leung et al., 2018; Macleod et al., 2015). Importantly, while we look to existing estimates to shape our expectations, the inherent heterogeneity between fields necessitates each field be assessed in turn.

Randomisation

Randomising group allocation reduces the risk of selection bias in experiments and evenly disperses confounders – known and unknown – between groups (Bebarta et al., 2003). Randomisation is essential for hypothesis-testing research and, without it, associated inferential statistics are invalid (Percie du Sert, Ahluwalia, et al., 2020).

A survey of systematic reviews of animal biomedical literature found a lack of randomisation correlated with increased effect sizes, demonstrating the impact of the absence of randomisation on the reliability of a field’s findings (Hirst et al., 2014). The failure to limit the influence of bias in preclinical research is found to undermine later translation attempts and results replications (Open Science Collaboration, 2015; Schmidt-Pogoda et al., 2020; Watzlawick et al., 2019).

Despite the implications of a lack of randomisation, reporting of this practice remains unsatisfactorily low in several preclinical fields (Table 3). A survey of biomedical preclinical research revealed just over one third of studies reported randomisation (Menke et al., 2020). A review of pain and anaesthesiology research saw 63% of articles reported randomisation (Fergusson et al., 2019). Due to some similarity in research areas, this rate may be more indicative of AMOA research. Further, as with masking, it appears rates of randomisation are increasing (Macleod et al., 2015). We hope to find similarly high rates of this measure in AMOA.

Table 3

Prevalence of Bias Minimisation Practices in Preclinical Research as Assessed by Previous Work

Study characteristic	Bebarta et al. (2008)	Kousholt et al. (2022)	Leung et al. (2018)*	Fergusson et al. (2019)	Hirst et al. (2014)	Ting et al. (2015)	Vesterinen et al. (2010)	Macleod et al. (2015)	Menke et al. (2020)
Field of interest	Emergency medicine	National survey (Denmark)	Animal welfare, analgesia or anaesthesia	Anaesthesiology, anaesthesia & analgesia, anaesthesia, British Journal of Anaesthesia	Biomedical	Rheumatology	Multiple sclerosis	8 Biomedical disease models	Biomedical
Publishing years of papers reviewed	1997-2001	2009 vs 2018	2009 vs 2015	2014-2016	1992-2012	2012	1961-2008	1992-2011	2018
Total number of paper	290	250 vs 250	236	282	31 systematic reviews	41	1152	2671	51 312
Masking (any mention)			19% vs 43%[a]						12.3%
Masked outcome assessment	10.7%	23.6% vs 38%		45%	35%	23.9%	16%	29.5%	
Masked allocation						15%			

Study characteristic	Bebarta et al. (2008)	Kousholt et al. (2022)	Leung et al. (2018)*	Fergusson et al. (2019)	Hirst et al. (2014)	Ting et al. (2015)	Vesterinen et al. (2010)	Macleod et al. (2015)	Menke et al. (2020)
Randomisation (any mention)	32.4%	24 vs 40.8[a]	50% vs 71% [a]	63%	29%	17.1%	9%	24.8%	36.3%
Sample size calculation		2.8% vs 12.8%	2.5% vs 10% [a]	29%[c]		0%	1%	0.7%	7.3%
Data exclusion (any mention)		20.4% vs 38.4%	65 v 67% [a]	37%		19.5%[b]			

Note. The first three rows describe the characteristics of the previous work. The following rows describe their findings.

[a] Only included studies where variable relevant

[b] This percentage describes reported attrition

[c] **T**here is ambiguity with this how the sample size calculation variable was coded. Despite efforts to contact the authors, it remains unclear. As such, we do not use this statistic in any comparisons

Sample size calculation

A study's sample size should be decided using a sample size calculation (SSC). This calculation involves an estimate of the expected effect size and an acceptable level of power. The expected effect size would ideally come from meta-analyses which aggregate effect sizes to avoid the influence of bias from a single study (Schäfer & Schwarz, 2019). This method ensures an appropriately powered study, a valid statistical model and trustworthy results (Flora, 2020; Percie du Sert, Ahluwalia, et al., 2020; Szucs & Ioannidis, 2017). Furthermore, hypothesis-testing research using inferential statistics must be adequately powered to certify the evidence being compared to the null hypothesis is suitably weighted (Percie du Sert, Ahluwalia, et al., 2020).

A literature built on well-powered studies is less likely to have false positives and inflated effect sizes and can thereby generate a more accurate understanding of an effect (Szucs & Ioannidis, 2017). An underpowered study, conversely, increases the chance of true effects being missed and effect sizes being overestimated (Landis et al., 2012; Macleod et al., 2008). Moreover, the pressure to achieve significant results combined with consistently underpowered research may increase the perceived necessity for researchers to engage in QRPs (Flora, 2020). When combined with publication bias, low power is associated with a decline in efficacy from the preclinical to the clinical stage (Schmidt-Pogoda et al., 2020).

Although possibly less of a concern, overpowered studies in animal research is unethical, as it places animals at unnecessary risk (Landis et al., 2012). This 'sweet spot' in sample size necessitates a power analysis in every study.

Research reveals many fields are chronically underpowered, including cognitive neuroscience, psychology, preclinical neuroscience, preclinical stroke, and preclinical multiple sclerosis (Button et al., 2013; Ellis, 2022; Fraley & Vazire, 2014; Schmidt-Pogoda et al., 2020; Szucs & Ioannidis, 2017; Vesterinen et al., 2010). Despite this, SSCs remain uncommon in preclinical research (Table 3). As such, it may be reasonable to expect similarly low levels in AMOA.

Data exclusion

Transparent reporting of data that are omitted from the final analyses is essential because of the implications for a study's power, the ability to accurately estimate effect sizes and the likelihood of finding a false positive (Miller, 2023).

This is especially true for research that uses small sample sizes, such as much preclinical animal research (Holman et al., 2016). In preclinical cancer and stroke research, non-reporting of excluded animals was associated with effect sizes that were likely overestimated (Holman et al., 2016).

Outlier exclusion may be one reason for data exclusion. It describes the practice of excluding data points or animals from analyses because they fall far from the mean. Outliers are suspected to be caused by a mechanism other than the one being studied, a malfunction of the apparatus or a spurious subject response (Cook et al., 2022; Simmons et al., 2011). While it seems beneficial to exclude irrelevant data, what is considered outlying is unstandardised in many research fields (Miller, 2023). This ambiguity presents an opportunity for bias to be introduced, as researchers may favour outlier definitions that lead to a significant result (Simmons et al., 2011).

The potential for bias associated with unreported data exclusion is exacerbated by other characteristics of the preclinical animal literature, which AMOA is unlikely to be immune to: publication bias, chronically low power, and underreporting of *a priori* inclusion and exclusion criteria (Holman et al., 2016; André, 2023). These factors combined can drastically increase the rate of false positives in a literature, increasing the risk that the findings may not reproduce or translate (Moher et al., 2016; Munafò et al., 2017).

Currently, preclinical animal research is yet to match the disclosure standards of clinical research in data exclusion reporting (Baker et al., 2014; Holman et al., 2016). The mixed prevalence estimates in the preclinical literature shown in Table 3 make it difficult to form a clear expectation for AMOA.

As well as these measures of transparency and reporting of bias minimisation practices, the current study examined two additional aspects that can directly affect the reliability of a study's results: the reporting of multiple comparisons adjustments (MCA) and the accuracy of reported statistical tests. As with all the variables studied here, these measures are relevant to most – if not all – research designs.

Multiple-comparisons adjustment

Statistical analysis often involves running multiple tests for a single hypothesis (Rubin, 2017). Doing so introduces the problem of multiplicity: with each additional test, the likelihood of finding a false positive. Using a statistical procedure to control for the multiple tests readjusts the false discovery rate (Gelman & Loken, 2013).

The necessity to adjust for multiple testing may be becoming increasingly important as data sets get larger and running analyses gets easier with improvements in technology and computational capacity (Leek & Storey, 2008; Niso et al., 2022). MCAs are necessary in a research field like AMOA that often uses large quantities of detailed data. For example, microarray studies looking for significant associations between an outcome and tens of thousands genetic details would, without adjustments, have an unacceptably high risk of finding false positives (Owzar et al., 2011).

This issue is exacerbated by underpowered research, together greatly undermining the reliability of results (Cramer et al., 2016; Gelman & Loken, 2013). Limited reliability reduces the stability and reproducibility of a finding which will have implications for translation potential (Khan et al., 2020; Lowenstein & Castro, 2009).

Despite the importance of MCAs, the reported prevalence is relatively understudied (Khan et al., 2020). The research that does exist is discouraging: a review of 819 psychology papers found that 47% used multiway ANOVA – where MCA is essential – but only 1% reported a correction procedure (Cramer et al., 2016). Estimates from cardiovascular and analgesic RCTs reported use of MCA in 28% and 45% of instances where it was required, respectively (Gewandter et al., 2014; Khan et al., 2020).

The current study will contribute to the sparse estimates of this practice.

Accurate reporting

Test statistic accuracy

The test statistics of a study report the type of test, the degrees of freedom, the test result, and the associated *p-value*. When the *p*-value is inconsistent with the associated test, the evidential value of the result is misrepresented (Nuijten & Polanin, 2020).

Given the reliance on null hypothesis significance testing in psychological research, it is essential that reported *p*-values are accurate (Flora, 2020; M. B. Nuijten et al., 2016). Inaccurate *p*-values contribute to the rate of false positives in the literature and therefore reduce its reproducibility (Nuijten & Polanin, 2020). In preclinical research, inaccurate *p*-values may influence a decision to pursue a treatment to clinical trials, placing humans at unnecessary risk for a potentially ineffective intervention.

Statistical inconsistencies may indicate 'sloppiness' in the research or review process, or engagement in QRPs (Green et al., 2018; Nuijten & Polanin, 2020). In a review of psychology researchers, about a fifth of respondents admitted to engaging in rounding down *p*-values, suggesting that inaccurate *p*-values are not always innocent mistakes (John et al., 2012). This supposition finds evidence in the fact that the inaccuracies found by Nuijten and colleagues (2016) were more often insignificant result incorrectly reported as significant. This is unsurprising given that researchers are incentivised to find significant results due to publication pressures (Giner-Sorolla, 2012).

To facilitate the detection of test statistic inconsistency, Epskamp and Nuijten (2015) developed statcheck. This R package recomputes *p*-values from the reported test and degrees of freedom and compares it to the published one. statcheck enabled Nuijten and colleagues (2016) to scan more than 30 000 papers for statistical inaccuracies. Discouragingly, Nuijten and colleagues (2016) found decisions errors – that is, inaccuracies that would change the significance of the statistical test at an alpha of .05 – in 13% of published psychology papers analysed. Similar findings in Canadian journals led the authors to recommend a statcheck or equivalent process to be included in the review process

(Green et al., 2018). Given the ease with which a study can be checked by such a program, published inconsistencies of this nature should almost never occur.

The limited estimates of test statistic accuracy mean any conjecture would be uninformed.

In estimating the prevalence of these variables, we aim to ascertain whether efforts to improve measures of reproducibility should include AMOA research. This may have implications for the field's translation potential.

In addition to these goals, we wondered whether certain attributes of AMOA research may influence the perception about its vulnerability to poor reproducibility.

Psychology's hierarchy of subdisciplines

Sciences are sometimes considered to range from 'hard' to 'soft' (Uher, 2021). However, beyond scientists' intuition, it is unclear what informs this hierarchy. While the natural sciences represent the harder sciences, psychology is considered soft (Fanelli, 2010). Researchers have tried to explain this intuition, suggesting the use of scientific methods or the level of noise in the data as explanations for the hierarchy (Fanelli, 2010; Uher, 2021). Using this logic, psychology may be considered soft as 'true experiments' are not always plausible, and humans are highly variant, complex units of study that often produce noisy data.

We suspect that a hierarchy of sorts may exist within psychology along similar lines. For example, subdomains that rely on experimental research design are considered harder than those that rely on observational designs. Animal behavioural research in psychology, for example, is considered harder than human behavioural research by some measures (Best et al., 2001; Kubina et al., 2008; Smith et al., 2000). A reason for this may be the increased level of intervention permitted in animal research and the greater ability to limit contextual influences. This may lead to less noise and larger effect sizes. Indeed, we suspected an aspect that informs this hierarchy within psychology may be the average magnitude of effects found in the subdisciplines of psychology.

Furthermore, we wondered if the replication crisis is considered more relevant to the softer psychology sub-disciplines that typically deal in smaller effect sizes, such as social and personality psychology, compared to harder sub-disciplines such as animal behaviour research.

To investigate the possibility that solutions to the replication crisis are relevant in subfields beyond those with small effect sizes, we firstly wanted to get an understanding of the average effect sizes in animal behavioural research, specifically in an addiction context. As such, we undertook a search for meta-analyses aggregating research looking at *in vivo* animal drug models (see Appendices A-D). We took 27 main effects from seven meta-analyses collectively analysing 200 papers. We found that, according to benchmarks, 20 effects would be classified as large, two medium and three small. This suggests that animal behavioural research does deal mainly in large effect sizes. This fact, along with the laboratory setting, the true experiment research design, and the ability of this field to develop invariant animal behavioural paradigms, may give reason for some to consider this field harder than others in psychology. Our next question, and the focus of this research paper, is whether efforts undertaken to restore credibility to some of psychology's subdisciplines following the reproducibility crisis are relevant to AMOA, one area of animal behavioural research? To answer this, we assessed the degree to which measures promoting transparency, accuracy and bias minimisation reporting are being implemented in this field.

Methods

This is a retrospective, observational study. It is exploratory and is the first of its kind in the animal addiction field. This means it is a discovery project aimed at uncovering the prevalence of the variables discussed in the introduction and presented in Table 4. As such, the results will be informative regardless of whether our expectations hold true.

This study was preregistered at https://osf.io/q2z4d/. Preregistration including deviations can be found in Appendices A and B.

Sample

Our sample process and exclusions can be seen in Figure 1. We used the search string below to find AMOA articles.

addict* OR substance abuse OR drug addiction OR drug treatment AND opioid OR opiate OR heroin AND treatment OR treat* AND behaviour* OR behavior*

We searched Scopus, Web of Science, PSYCinfo and PubMed. We limited results to "article" or "empirical study", to "animal", written in English and published between 2019 and 2023. In Scopus, results were also limited to relevant research areas (neuroscience, psychology, pharmacology, toxicology, and pharmaceutics, and multidisciplinary).

For a study to be included it had to satisfy the following criteria: be an empirical study; include some *in vivo* study of animals; include some testing of opioids; study the effects of opioids, a treatment of opioid addiction or alternatives to opioid analgesics. The exclusion criterion was that the article had not been published in a journal.

We initially expected to take a random sample of the AMOA literature. However, upon completing the search we found that the number of papers located was feasible and so was taken in its entirety.

Figure 1

PRISMA Flow Chart of Animal Models of Addiction Search

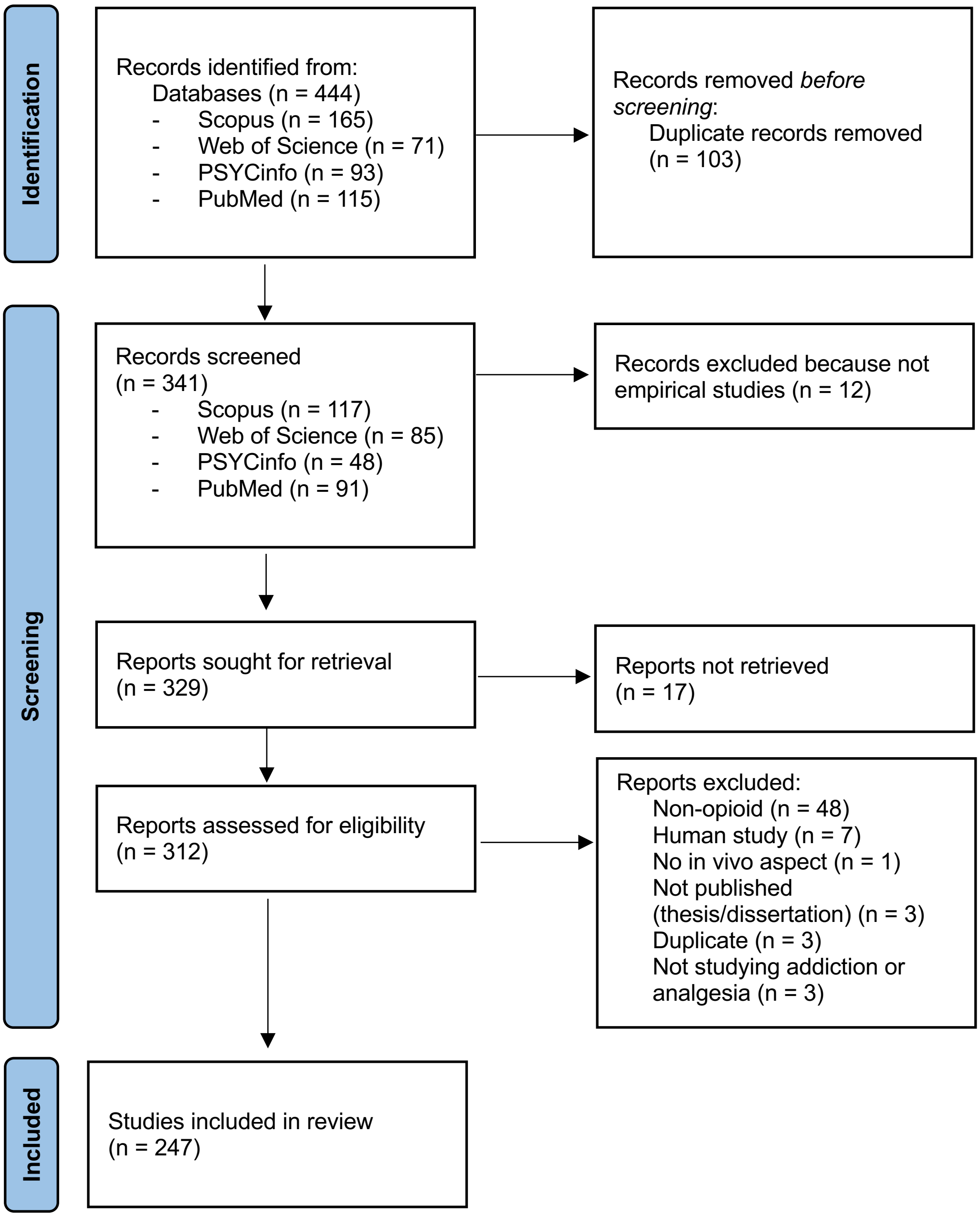
Identification
Records identified from:
Databases (n = 444)
- Scopus (n = 165)
- Web of Science (n = 71)
- PSYCinfo (n = 93)
- PubMed (n = 115)
Records removed *before screening*:
Duplicate records removed (n = 103)
Screening
Records screened (n = 341)
- Scopus (n = 117)
- Web of Science (n = 85)
- PSYCinfo (n = 48)
- PubMed (n = 91)
Records excluded because not empirical studies (n = 12)
Reports sought for retrieval (n = 329)
Reports not retrieved (n = 17)
Reports assessed for eligibility (n = 312)
Reports excluded:
Non-opioid (n = 48)
Human study (n = 7)
No in vivo aspect (n = 1)
Not published (thesis/dissertation) (n = 3)
Duplicate (n = 3)
Not studying addiction or analgesia (n = 3)
Included
Studies included in review (n = 247)

Table 4

Study Characteristics Assessed in the Current Study

Study Characteristic	Response options	ARRIVE 2.0 guideline (where applicable)	Search terms & any additional instructions
Original or replication	Original Replication Unsure		Read abstract *Replicat*
Preregistration	No statement of preregistration Yes, statement of preregistration with link Yes, statement of preregistration but no link There is a statement of non-preregistration This paper is a registered report	19. Provide a statement indicating whether a protocol (including the research question, key design features, and analysis plan) was prepared before the study, and if and where this protocol was registered.	*Regist, osf, aspredicted, preclinicaltrials*
Data availability	No statement regarding data availability Yes, statement that some raw data is available via link Yes, statement some data available but link broken	20. Provide a statement describing if and where study data are available.	*Availab, request, reposit, data*

Study Characteristic	Response options	ARRIVE 2.0 guideline (where applicable)	Search terms & any additional instructions
	Yes, statement some data available but link absent Unavailable - statement that the data is unavailable Upon request		
Code availability	No code or syntax for analysis available Yes, syntax/code provided Upon request		*Code, syntax, script*
ARRIVE	Yes, statement of compliance with ARRIVE or ARRIVE checklist in supplementary materials No mention of ARRIVE or compliance with another set of reporting guidelines Other - Mention of compliance with other guidelines		*Arrive, guide, accordance, protocol, reporting*
Masking	Yes, blinding mentioned in relation to this study	5. Describe who was aware of the group allocation at the different stages of the experiment (during the	*Blind, mask*

Study Characteristic	Response options	ARRIVE 2.0 guideline (where applicable)	Search terms & any additional instructions
	No blinding mentioned in relation to this study Statement of no blinding/masking used	allocation, the conduct of the experiment, the outcome assessment, and the data analysis).	
Randomisation	Yes, randomisation mentioned in relation to this study Other method of group allocation given No allocation method mentioned Statement of NO randomisation	4 a. State whether randomisation was used to allocate experimental units to control and treatment groups.	*Random, alloc, assign*
Sample size justification	No justification given Power analysis/sample size planning Past research Practical constraints	2b. Explain how the sample size was decided. Provide details of any *a priori* sample size calculation, if done. *If you have used an a priori sample size calculation, report* • *the analysis method (e.g., two-tailed Student t test with a 0.05 significance threshold)* • *the effect size of interest and a justification explaining why an effect size of that magnitude is relevant* • *the estimate of variability used (e.g., standard deviation) and how it was estimated*	Read section 'subjects' or 'animals' in methods, *Power, plan, priori*

Study Characteristic	Response options	ARRIVE 2.0 guideline (where applicable)	Search terms & any additional instructions
		• the power selected	
Multiple corrections	Corrected No mention of correction method	7a. Provide details of the statistical methods used for each analysis, including software used. *Relevant information to describe the statistical methods include:* *• the outcome measures* *• the independent variables of interest* *• the nuisance variables taken into account in each statistical test (e.g. as blocking factors or covariates),* *• what statistical analyses were performed and references for the methods used* *• how missing values were handled* *• adjustment for multiple comparisons* *the software package and version used, including computer code if available*	*Correct, bonf, holm, scheffe, tukey, Benj, family, FDR, false*
Exclusion	No statement of animal exclusion Yes, animals were excluded from the study Statement of no animal exclusion	3 b. For each experimental group, report any animals, experimental units, or data points not included in the analysis and explain why. If there were no exclusions, state so.	*Exclu, outl, discard, sacrif*

Study Characteristic	Response options	ARRIVE 2.0 guideline (where applicable)	Search terms & any additional instructions
Exclusion reasons	Outlier exclusion Other Outlier exclusion AND other reason(s) No reason given Not applicable (no exclusion mentioned)		
Supplementary files [a]	Yes No Yes, but link absent/broken		*Supplementary, supporting, appendi*

Note. Text in italics taken from elaborated version of guidelines (Percie du Sert, Ahluwalia, et al., 2020).

[a] Serves to remind coders to check supplementary files. Not a variable of interest.

Pilot coding

All articles were coded by two coders, which is the gold standard for this research design. Coders used a Google sheet codebook (see Table 4 for variables and response options; see https://osf.io/q2z4d for Excel version) developed by the four coders throughout the pilot coding process. After included variables had been finalised, Coder 1 completed the first round of pilot coding on five articles. After Coder 1 had ensured functionality of the codebook, Coder 1, Coder 2 and Coder 3 coded a small selection of studies together and further adjusted the codebook. Next, all four coders were given 5 articles to code. Coding proper was commenced when all four coders agreed on the responses and were satisfied that the search terms were effective (Table 4).

All pilot coding articles were selected from a search of preclinical addiction literature not specific to opioids. The wider pool of articles meant a low likelihood of overlap between the pilot coding sample and the final sample.

Coding procedure

The coding instructions (Appendix E) were developed to standardise our coding procedure. In essence, each variable's relevant search terms (Table 4) were looked for in the article using the search function. Some variables also required scanning relevant parts of the article. For example, to detect sample size justification, the 'subjects' or 'animals' paragraph was read, and search terms were looked for. 90.6% of articles were double coded.

The average percentage agreement of responses was 91.6% (range: 89.43%-92.87%) and average Kripendorff's alpha was .93 (range: .74-1) (Tables 5 and 6).

Between 51 and 91 articles were allocated to each coders Coder 2, Coder 3 and Coder 4, and all articles were double coded by Coder 1. An additional 23 papers were not double coded due to the time constraints of one coder. However, given the high level of interrater reliability, we do not consider this a major limitation.

Importantly, coders were instructed to code generously – that is, we wanted to be biased in the charitable direction. This was in recognition of the fact that our measures are imperfect, and we may risk systematically underestimating the prevalence of some practices by not including a relevant search term. We did not want this to detract from our main goal of discerning whether attempts to improve reproducibility are relevant to AMOA. By coding in a manner that gives the benefit of the doubt, we hope to partly compensate for any threats to the accuracy of our estimates in this respect. As such, we aimed to estimate the upper bound of the prevalence of each variable.

Where supplementary files were available, these were checked for all characteristics but were not scanned for test statistics. A variable for the presence or absence of supplementary files was created to remind the coders to check it, but this was not a variable of interest.

Table 5

Percentage Agreement Between Coders

Variable	Coder 1 and Coder 2	Coder 1 and Coder 3	Coder 1 and Coder 4	All coders
% agreement	92.5%	92.9%	89.4%	91.6%

Table 6

Kripendorff's Alpha Interrater Reliability for All Variables

Variable	Coder 1 and Coder 2	Coder 1 and Coder 3	Coder 1 and Coder 4	All coders
Original or replication	1	1	0.99	.99
ARRIVE guidelines	0.88	0.85	0.74	.82
Preregistration	0.99	0.90	1	.96
Supplementary files	0.97	0.83	0.97	.92
Masking	0.97	0.84	0.98	.93
Randomisation	0.92	0.79	0.92	.91
Sample Size Justification	1	1	0.98	

Variable	Coder 1 and Coder 2	Coder 1 and Coder 3	Coder 1 and Coder 4	All coders
Reason for expected effect size	1	1	1	1
Expected effect size type	1	1	1	1
Expected effect size	0.98	1	1	
Multiple corrections	0.90	0.81	0.86	.86
Exclusion	0.97	0.85	0.92	.93
Reason for exclusion	0.94	0.93	0.92	.93
Number of *p*-values	0.85	0.89	0.96	.9
Number of non-decision errors	0.84	0.82	0.97	.88
Number of decision errors	0.86	0.88	0.88	.87
Average	0.94	0.90	0.94	0.93

Note. These calculations were run on SPSS.

Statistical Analyses

Our results are largely descriptive: we are interested the proportions of articles that satisfied each target variable. Unless otherwise specified, the denominator is the total sample size (247). We also report the 95% confidence intervals based on the adjusted Wald interval (Bonett & Price, 2012).

Additionally, we compared our results to those of previous research using a two-sample proportions z-test. This calculation tests for significant differences between two proportions. We used the results from previous studies as benchmarks, meaning we did not account for the uncertainty of their estimates. All z-tests were one-tailed.

These analyses were run on SPSS and online calculators that used the pwr library in R (Sauro, 2023; Statskingdom, 2022b).

Results

Our final sample consisted of 247 articles published between 2019 and 2023. After data collection had finished, Coder 1 verified any discrepancies between coders in ways described in Appendix F. In essence, responses that estimated the upper bound of prevalence rates were favoured and the opinion of Coder 4 was sought to settle ambiguous cases.

Sample characteristics

The sample characteristics can be found in Tables 7 and 8. Table 7 shows the frequencies of the years the articles were published. Table 8 presents the journals that contained more than two articles in our sample (25 in total) as well as whether they endorse the ARRIVE guidelines and provide the registered report format. A list of all journals is provided in Appendix G.

Table 7

Number of Animal Models of Opioid Addiction Articles Published Per Year

Year	Number of articles	Percent
2019	46	18.6%
2020	60	24.3%
2021	62	25.1%
2022	56	22.7%
2023	23	9.3%
Total	247	100%

Table 8

Top Journals in Animal Models of Opioid Addiction Sample

Journal name	Frequency	Percent	ARRIVE endorsement	Accept registered reports
Neuropharmacology	17	6.9	Y	N
Addiction Biology	15	6.1	N	N
Neuropsychopharmacology	11	4.5	N	N
Psychopharmacology	10	4.0	N	N
Behavioural Brain Research	8	3.2	Y	N
Frontiers In Pharmacology	8	3.2	N	N
Drug And Alcohol Dependence	7	2.8	Y	Y
Neuroscience Letters	7	2.8	N	N
International Journal of Molecular Science	6	2.4	Y	N
Frontiers In Molecular Neuroscience	5	2.0	N	N
Journal Of Pharmacology And Experimental Therapeutics	5	2.0	N	N
Pharmacology Biochemistry And Behavior	5	2.0	Y	N
Addiction Biology	4	1.6	N	N
Frontiers In Behavioral Neuroscience	4	1.6	N	N
Journal of Neuroscience	4	1.6	N	N
Journal Of Psychopharmacology	4	1.6	N	N
Pharmacology, Biochemistry And Behavior	4	1.6	N	N
Acta Pharmacologica Sinica	3	1.2	Y	N
Behavioural Pharmacology	3	1.2	N	N
Frontiers In Neuroscience	3	1.2	N	Y
International Journal Of Neuropsychopharmacology	3	1.2	N	N
Molecular Psychiatry	3	1.2	N	N
Pain	3	1.2	N	N
Progress In Neuro-Psychopharmacology & Biological Psychiatry	3	1.2	Y	N
Translational Psychiatry	3	1.2	N	N
Total	148	59.5	7[a]	2[a]

Note. Only journals with more than 2 articles were included in this table. For the full list of journals, see Appendix G. ARRIVE endorsement as reflected on the ARRIVE website (ARRIVE Guidelines, 2023). Registered report adoption as reflected on the Centre for Open Science: Registered Reports website (Centre for Open Science, 2023).

[a] Number of 'Y' responses

Transparency and Replication

Overall, transparency measures were not common in AMOA. The prevalence of these variables can be found in Table 9. The results of all two-sample proportion *z*-tests can be found in Table 10.

Replications

We found no replications (0%, 95% CI [0, 1.3]).

Preregistration

There were no preregistered articles (0%, 95% CI [0, 1.3]). This finding was significantly less than the (73%) found in smoking RCTs by Norris and colleagues (2021), Z=10.72, p<.001.

Open data

Available data was found in 8 (3.2%, 95% CI [1.5, 6.4]) articles, while 59 (23.9%, 95 CI [19, 29.6]) papers contained 'available upon request' statements. This left 174 (70.4%, 95% CI [64.5, 75.8]) papers that made no mention of raw data availability which was a significantly smaller amount than previous estimates: (Adewumi and colleagues (2021) (87%), Z=2.87, p=.004; Hamilton and colleagues (2023) (92%), Z=3.91, p<.001; Hardwicke and colleagues (2020) (92.3%), Z=3.98, p<.001.

Open code

There were no (0%, 95% CI [0, 1.3]) articles shared their code but 2 (.8%, 95% CI [.03, 3.1]) had 'available upon request' statements.

Table 9

Results from the Current Study

Study characteristic		Results % [95% CI]	Results (n)
Replication	Original	100% [96.7, 100]	246
	Replication	0% [0, 1.3]	0
	Unsure	0% [0, 1.3]	1
Preregistration	States preregistered with link	0% [0, 1.3]	0
	States preregistered with no link	0% [0, 1.3]	0
	States not preregistered	.4% [0, 0.02]	1
	No statement	99.6% [97.5, 100]	246
	Registered report	0% [0, 1.3]	0
Data availability	States available and accessible	3.2% [1.5, 6.4]	8
	States available but link broken	.4% [0.01, 2.5]	1
	States available but link absent	2% [0.7, 4.8]	5
	Data unavailable	0% [0, 1.3]	0
	States available upon request	23.9% [19, 29.6]	59
	No statement	70.4% [64.5, 75.8]	174
Code availability	States available	0% [0, 1.3]	0
	States available upon request	.8% [.03, 3.1]	2
	No statement	99.2% [96.9, 99.9]	245
Supplementary information	Yes	44.1% [38.1, 50.4]	109
	Yes, but link absent or broken	5.3% [3, 8.8]	13
	No	50.6% [44.4, 56.8]	125
Total		100%	247

Table 10

Results from Two-Sample Proportion Z-Tests Comparing the Current Study's Findings to Previous Findings of Estimates of Reproducibility and Transparency Practices

Study characteristic	Our result	Comparative study	Comparative result	Z-statistic	P-value of comparison[a]
Replications	0%	Adewumi et al., 2021	0.4%	0.63	.263
		Piu et al., 2022	0.2%	0.45	.327
Preregistration					
Statement of preregistration	0%	Hardwicke et al., 2020	0%	-	-
		Adewumi et al., 2021	2.9%	1.74	.083
		Norris et al., 2021	73%	10.72	<.001
Statement of no preregistration	.4%	Adewumi et al., 2021	0%	0.63	.263
		Hardwicke et al., 2020	0%	0.63	.263
Data availability					
No statement	70.4%	Adewumi et al., 2021	87%	2.87	.004
		Hamilton et al., 2023	92%	3.91	<.001
		Hardwicke et al., 2020	92.3%	3.98	<.001

Study characteristic	Our result	Comparative study	Comparative result	Z-statistic	P-value of comparison[a]
Statement of availability and accessible	3.2%	Hamilton et al., 2023	2%	0.53	.703
		Adewumi et al., 2021	8.2%	1.52	.064
Code availability					
No statement	99.2%	Hardwicke et al., 2020	98.7%	0.35	.636
		Norris et al., 2021	99%	0.15	.560
		Adewumi et al., 2021	99.2%	0	.500
		Hamilton et al., 2023	99.5%	0.26	.396

Note. These comparisons were done using an online two-sample proportions *z*-test calculator that used the pwr library in R (Statskingdom, 2022a). Alpha is set at *p*=.05 and all tests were one-tailed. In previous studies where two time periods were sampled, the proportion from the most recent time period is used (see Table 2 for further details).

[a] p<.05 indicates a statistically significant difference between the current study's estimate and a previous study's estimate.

Reporting Standards

Table 11 presents all proportions of these variables. Table 12 compares our results to the results of other works using the two-sample proportions *z*-test.

Masking

Some mention of masking was made in 88 (35.6%, 95% CI [29.9, 41.8]) papers which was significantly more than Menke and colleagues (2020) (12.3%), Z=3.86, p=.001 but not significantly different Leung and colleagues (2018) (43%), Z=1.07, p=.142.

Randomisation

Randomisation was mentioned in 120 (48.6%, 95% CI [42.4, 54.8]) papers which was significantly more than the result of Menke and colleagues (2020) (36.3%), Z=1.76, p=.039, and not significantly different to the results of Kousholt and colleagues (2022)(40.8%), Z=1.11, p=1.34 or Fergusson and colleagues (2019) (63%), Z=2.05, p=.020. However, it was significantly less than the estimate of Leung and colleagues (2018) (71%), Z=3.23, p<.001.

Sample size calculation

12 (4.9%, 95% CI [2.7, 8.4]) papers reported using a power analysis or sample size planning to justify their sample size. This proportion was not significantly different to the estimates found by Menke and colleagues (2020) (7.3%), Z=0.71, p=.239 or Leung and colleagues (2018) (10%), Z=1.37, p=.085. Although it was significantly less than the estimate of Kousholt and colleagues (2022) (12.8%), Z=1.97, p=.025. A further breakdown of types of sample size calculations can be found in Table 13.

Data exclusion

80 (32.4%, 95% CI [26.9, 38.5]) articles reported excluding data from the study, which was not significantly less than estimates from Fergusson and colleagues (2019) (37%), Z=0.68, p=.247 or Kousholt and colleagues (2022) (38.4%, Z=0.89, p=.187. The estimate of Leung and colleagues (2018), however, was significantly larger (67%), Z=4.89, p<.001. Of the 80 articles that reported excluding animals or data, 9 (11.25%, 95% CI [5.8,

20.2]) reported outlier exclusion and 7 (8.75%, 95% CI [4, 17.2]) reported outlier exclusion and another reason, meaning a total of 16 (20%, 95% CI [12.6, 30.2]) papers reported some outlier exclusion. An additional 65 (81.25%, 95% CI [71.2, 88.4]) papers reported another reason for excluding animals or data, meaning a total of 72 papers excluded some data for a reason other than outlier exclusion. This information is presented in Table 14.

Masking, randomisation, sample size justification and exclusion

Only 4 (1.6%, 95% CI [.05, 4.2]) papers contained some mention of all four bias minimisation measures.

Multiple-comparisons adjustment

Of the 247 papers in the sample, 189 (76.5%, 95% CI [70.8, 81.4] reported a multiple comparisons adjustment which was a significantly greater proportion than those found by Khan and colleagues (2020) (28.3%), Z=6.87, p<.001 and Gewandter and colleagues (2014) (45%), Z=4.56, p<.001.

Table 11

Results of Reporting Variables from the Current Study

Study Characteristic		Results % [95% CI]	Results (n)
ARRIVE	Statement of compliance	7.3% [5, 11]	18
	Other reporting guidelines followed	0% [0, 1.3]	0
	No statement of reporting guidelines	92.7% [88.7, 95.4]	229
Masking	Yes, mentioned	35.6% [29.9, 41.8]	88
	Statement of no masking	1.6% [.4, 4.2]	4
	No mention	62.8% [56.6, 68.6]	155
Randomisation	Yes, mentioned	48.6% [42.4, 54.8]	120
	Other allocation method mentioned	6.5% [4, 10.3]	16
	Statement of no randomisation	.8% [.03, 3.1]	2
	No mention	44.1% [38.1, 50.4]	109
Sample size justification	Power analysis/sample size planning	4.9% [2.7, 8.4]	12
	Past research	.8% [.03, 3.1]	2
	Practical constraints	0% [0, 1.3]	0
	No justification	94.3% [90.6, 96.7]	233
Exclusion	Animal or data excluded	32.4% [26.9, 38.5]	80
	Statement of no exclusion	3.2% [1.5, 6.4]	8
	No statement	64.4% [58.2, 70.1]	159
Multiple corrections	Mention of correction	76.5% [70.8, 81.4]	189
	No mention of correction	23.5% [18.6, 29.2]	58
Total		100%	247

Table 12

Results from Two-Sample Proportion Z-Tests Comparing the Current Study's Findings to Previous Findings of Estimates of Bias Minimisation and Accurate Reporting Practices

Study characteristic	Our result	Comparative study	Comparative result	Z-statistic	P-value of comparison[a]
Masking (any mention)	35.6%	Menke et al., 2020	12.3%	3.86	.001
		Leung et al., 2018	43%[c]	1.07	.142
Randomisation (any mention)	48.6%	Menke et al., 2020	36.3%	1.76	.039
		Kousholt et al., 2020	40.8%	1.11	.134
		Fergusson et al., 2019	63%	2.05	.020
		Leung et al., 2018	71%	3.23	<.001
Sample size calculation [b]	4.9%	Menke et al., 2020	7.3%	0.71	.239
		Leung et al., 2018	10%	1.37	.085
		Kousholt et al., 2022	12.8%	1.97	.025
Data exclusion (any mention)	32.4%	Fergusson et al., 2019	37%	0.68	.247
		Kousholt et al., 2022	38.4%	0.89	.187

Study characteristic	Our result	Comparative study	Comparative result	Z-statistic	P-value of comparison[a]
Multiple comparisons corrections present	76.5%	Leung et al., 2018	67%	4.89	<.001
		Khan et al., 2020	28.3%	6.87	<.001
		Gewandter et al., 2014	45%	4.56	<.001

Note. These comparisons were done using an online two-sample proportions *z*-test calculator that used the pwr library in R (Statskingdom, 2022a). Alpha is set at *p*=.05 and all tests were one-tailed. In previous studies where two time periods were sampled, the proportion from the most recent time period is used (see Table 3 for further details).

[a] p<.05 indicates a statistically significant difference between the current study's estimate and a previous study's estimate.

[b] This variable reflects the proportion of studies that used power analyses or other sample size planning calculations used. This variable was measured in our study as the 'power analysis/sample size planning' response option of 'sample size justification'.

Table 13

Results Breakdown for Articles Including Sample Size Calculations

Study Characteristic		Results % [95% CI]	Results (n)
Reason for expected effect size	Past research	33.3% [13.6, 61.2]	4
	No reason provided	66.3% [38.8, 86.5]	8
Effect size type	Not mentioned	100% [78.4, 100]	12
Expected effect size	0.5	8.3% [.01, 37.5]	1
	0.5-0.9	8.3% [.01, 37.5]	1
	Not mentioned	83.3% [54, 96.5]	10
Total		100%	12

Note. This table presents a further breakdown of the results from articles that used sample size calculations (power calculations or other sample size planning techniques) to calculate the study's sample size.

Table 14

Results Breakdown for Articles Including Mention of Data Exclusion

Study Characteristic		Results % [95% CI]	Results (n)
Exclusion reasons	Outlier	11.3% [5.8, 20.2]	9
	Outlier and other	8.8% [4, 17.2]	7
	Other	81.3% [71.2, 88.4]	65
Total		100%	80

Note. This table presents a further breakdown of the results from articles that mentioned exclusion of data or animals. Percentages may not total 100 due to rounding.

Accurate Reporting

Table 15 presents the proportions of the following results. Table 16 compares these results to the findings of Nuijten and colleagues (2016).

Detection rate

Test statistics were detected in 185 (74.9%, 95% CI [69.1, 79.9]) papers which was significantly more than Nuijten and colleagues (2016) result of 54.4%, Z=3.03, *p*=.001.

Non-decision errors

Non-decision errors were detected in 96 (51.9%, 95% CI [44.7, 59] papers in our sample which was not statistically larger than Nuijten and colleagues (2016) result (49.6%), Z=0.33, *p*=.372.

Decision errors

Decision errors were found in 21 papers (11.4%, 95% CI [7.5, 16.8]) which was not significantly smaller than Nuijten and colleagues' (2016) result (12.9%), Z=0.32, p=.373.

Table 15

Test Statistic Results

Study Characteristic		Results (n)	Results % [95% CI]
Test statistics	Papers with test statistics detected	185	100%
	Total statistics detected	5144	
Non-decision errors	Papers with non-decision errors detected	96	51.9% [44.7, 59]
	Total errors detected	302	
Decision errors	Papers with decision errors	21	11.4% [7.5, 16.8]
	Total decision errors detected	43	

Note. Percentages reflect the proportion of articles out of the 185 papers where any test statistics were detected.

Table 16

Test Statistic Results from Two Sample Proportion Z Tests Comparing the Current Study's Findings to Previous Findings

Study characteristic	Our result	Comparative study	Comparative result	Z-statistic	P-value of comparison[a]
Test statistic detection rate	74.9%	Nuijten et al., 2016	54.4%	3.03	.001
Percentage of papers containing a non-decision error	51.9% [b]	Nuijten et al., 2016	49.6%	0.33	.372
Percentage of papers containing a decision error	11.4% [b]	Nuijten et al., 2016	12.9%	0.32	.373

Note. These comparisons were done using an online two-sample proportions *z*-test calculator that used the pwr library in R (Statskingdom, 2022a). Alpha is set at *p*=.05 and all tests were one-tailed.

[a] p<.05 indicates a statistically significant difference between the current study's estimate and a previous study's estimate.

[b] This percentage includes only those papers where test statistics were detected, N = 185.

Discussion

Reproducibility in preclinical research is supported by transparent research and reporting that is thorough and accurate (Munafò et al., 2017). Research generated by a field with good reproducibility may have better translation potential (Fergusson et al., 2019; Landis et al., 2012).

This study investigated the prevalence of transparency and thorough and accurate disclosure practices in the AMOA. This was to determine to what extent such measures are already in use, and if recent efforts to improve reproducibility and translation rates may be relevant to.

In the first study of its kind in the AMOA literature, we manually reviewed papers studying opioid use and opioid alternatives to characterise if they fulfilled the target variables.

When interpreting this study's results, it is important to note that we are estimating the rate at which these practices are reported. This is not necessarily the same as how often they are implemented. However, evidence shows that research with poor reporting of bias minimisation practices is associated with overestimates of effects sizes (Bebarta et al., 2003; Crossley et al., 2008; M. R. Macleod et al., 2008; Riley et al., 2016; Rooke et al., 2011; Tikka et al., 2021; Vesterinen et al., 2010). These studies concluded that this inflation was likely caused by bias in the research design, probably introduced by the absence of the bias minimisation measures that were not reported. This suggests a lack of reporting may indeed reflect a lack of doing.

Furthermore, science is based on transparency and verifiability (Munafò et al., 2017). A consumer of science should not have to trust that a certain practice was implemented; it should be clearly stated. As such, the sceptical reader will assume a procedure was omitted from the experiment if there is no mention in the report. While this leaves some room for ambiguity, we believe it is fair to judge an experiment based on its report. Therefore, we believe our interpretations are fair and justified.

Lastly, we encourage caution when interpreting the comparisons to previous studies. Due to the novelty of this study, we do not have estimates from more closely associated fields. Where possible, we attempted to minimise the numerous differences between the compared fields by favouring more recent estimates from studies with some commonality to AMOA.

Replication

There were no replications in our sample of the AMOA literature. This finding was not different to the rates of replications found in addiction research or psychology research (Adewumi et al., 2021; Pui Yu Lee, 2022). This was in keeping with our expectations of low replication rates as informed by estimates from the fields of addiction and psychology (Makel et al., 2012; Norris et al., 2021). While not unexpected, this result is informative as it is the first to indicate that replication rates in AMOA are comparable to associated fields.

There are several possible interpretations of this result. The first is that replications are not being done in AMOA, possibly due to the pressures of working in a competitive field that rewards novelty over replications (Gorman, 2019).

Another interpretation is that replications are being done, but they are not being published. For example, a researcher may replicate a foundational effect in a preliminary experiment before building on it. Publishing space limitations, however, may prohibit this replication from being published with the novel experiment.

If this is the case, there are solutions: sharing of data and results on preclinical registries is free and straightforward. This practice also combats research waste and facilitates more accurate estimates of effects in data aggregation efforts (Chin, 2023; Moher et al., 2016; van der Naald et al., 2020).

Considering this study’s findings of low rates of bias minimisation practices and a lack of engagement in transparency practices, the absence of reported replications may be particularly concerning. This result adds weight to calls for replication attempts in the field of addiction, including AMOA (Heirene, 2021).

Transparency

Preregistration

Our sample contained no articles that were preregistered, one that had a statement of non-preregistration and no registered reports. Our result was smaller by a substantial amount than the rate of preregistration found in clinical addiction research, but not different to prevalence estimates in the social sciences and addiction broadly (Adewumi et al., 2021; Hardwicke et al., 2020; Norris et al., 2021). Despite our suggestion that working alongside clinical research would encourage preregistration in the AMOA field, it appears that current preregistration habits are more in line with psychology and the social sciences than they are with clinical addiction research.

The recent creation of animal-specific registries Preclinical Trials and Animal Study Registry has addressed concerns that such an absence was one reason for the low rates of preclinical preregistration (Ting et al., 2015; van der Naald et al., 2021). However, the number of studies preregistered on these platforms remains discouragingly low (van der Naald et al., 2021).

A lack of awareness about preregistration and its associated benefits may remain an obstacle for AMOA researchers, as it is in other fields (Percie du Sert, Hurst, et al., 2020; van der Naald et al., 2021). Alternatively, investigators working in AMOA may find the additional work required to preregister burdensome or they may be unwilling to preregister due to a desire to safeguard their intellectual property (Kimmelman & Anderson, 2012; Nosek et al., 2019).

Proponents of preregistration would argue that these obstacles can be overcome with improved instruction. Firstly, researchers may be more inclined to make the additional effort if they are aware of the value of preregistration in reducing research waste, facilitating detection of QRPs and HARKing, and minimising the impact of publication bias (Nosek et al., 2018; Percie du Sert, Hurst, et al., 2020).

Secondly, animal registries continue to try to streamline the preregistration process to make it more efficient and easier to use for researchers new to the practice (van der Naald et al., 2021). Lastly, the option to embargo a preregistration is available to help assuage concerns about intellectual property theft (van der Naald et al., 2021)

Slow uptake of the registered report format may be influenced by many of the same obstacles as preregistration as well as the apparently low number of participating journals publishing AMOA research as seen in Table 8.

Critics of preregistration may say that its importance has been overstated (Devezer et al., 2021; Rubin, 2017). We maintain, though, that in the absence of 'contemporary' transparency, it is one part of a multi-pronged solution to address the practices and systemic influences that have led some fields to crisis point (Rubin, 2017). AMOA would benefit from higher rates of preregistration, as it encourages researchers to consider the use of bias minimisation techniques and power analyses. As we have found, AMOA has room for improvement in these domains.

Open data

Overall, this study found low rates of data sharing. The large majority (70.4%) made no mention of data availability. However, this proportion was significantly lower than the equivalent in addiction, social sciences, and a large-scale review of preclinical and clinical health and medicine metascience (Adewumi et al., 2021; Hamilton et al., 2023; Hardwicke et al., 2020). This reveals a relatively good awareness of data sharing as a practice, or the considerable number of journals that require a data availability statement. We consider both possibilities promising.

Less promising was the 3.2% of articles that had accessible data. This proportion was not different to estimates in addiction and health and medicine metaresearch (Adewumi et al., 2021; Hamilton et al., 2023). This contradicts our anticipation that AMOA researchers may engage in this practice relatively frequently because of the existing familiarity with data sharing practices in the form of data repositories (Munafò, 2015).

We had also hoped this familiarity would mean those working in AMOA research would be aware of the superiority of online databases for storage, leading to low instances of 'data available upon request' statements. However, almost one quarter of all articles and

80% of articles with any data statement were ‘available upon request’ statements. Hardwicke and Ioannidis (2018) revealed the inadequacy of this data sharing solution when they were unable to retrieve 68% of study data from authors post-publication.

Interestingly, coders came across several data statements that suggested a misunderstanding of ‘data availability’ as availability of analysed data. This misinterpretation may be behind statements that the ‘data are contained within the article’. A similar data availability statement template can be found on the Taylor & Francis website: ‘data are contained within the article [and/or] its supplementary materials’ (Talylor & Francis, 2023). This is understandably confusing.

While it is reasonable that the raw data may be in the supplementary files, we suspect it is often implausible to present raw data in the article. One possible exception is if the raw data is presented in a graph from which can be extracted using a data extraction tool (WebPlotDigitizer, 2022)**.** However, if the goal is the efficient sharing of accurate raw data, this method may not be ideal.

Similarly, we found data availability statements that were unclear or unaccompanied by a link or further description about how to access the data. This, as well as the misinterpretation of data availability, suggest a lack of involvement on behalf of the journal in verifying meaningful compliance with open data policies. Such involvement is imperative for improving rates of open practices (Hair et al., 2019).

Lastly, coders encountered statements that said data would be shared after a period of embargo. We consider this a positive, as it indicates researchers are finding ways to share data that do not conflict with their other interests. We hope, though, that clinicians wanting to pursue preclinical treatments are excluded from such an embargo, and that embargoed data is uploaded to an appropriate repository upon publication to avoid similar issues to the ‘upon request’ method.

The low rates of data availability and preregistration would likely preclude interested parties from assessing rates of research non-publication and underreporting of animals used in AMOA research (van der Naald et al., 2020). Because of the implications for informed decision making at the clinical stage and the efficient use of funding and animals, understanding the rates of field's non-publication and underreporting is essential.

While actual data availability remains low, we consider our results encouraging for future improvement. Improving data sharing will require attentive participation on behalf of journals, and adequate funding from relevant bodies to allow for the additional time this practice may take (Munafò et al., 2017).

Open code

We found no instances where code was available, and two instances where it was available 'upon request'. The proportion of studies that made no mention of code availability were not different to estimates in the fields of addiction, clinical addiction, social sciences and health and medicine metaresearch (Adewumi et al., 2021; Hamilton et al., 2023; Hardwicke et al., 2020; Norris et al., 2021). This places AMOA on par with a range of research fields.

Several potential roadblocks to the wider adoption of code sharing have been proposed, ranging from the practice (time, adequate funding, lack of know-how) to concerns about potential misuse or misinterpretation of code and the data it analyses (Naudet et al., 2018).

Improving know-how will require training in code sharing procedures. This could take the form of practical modules for researchers, although the motivation to engage in additional instruction may need to come from policies by journals and funders (Munafò et al., 2017). Gomes and colleagues (2022) believe the potential for the misuse of analysis code can also be addressed with education on how to include all relevant information, such as assumptions and caveats. Crucially, having appropriate time and funding to dedicate to preparing the code and data for sharing underlines the necessary involvement of funders in the adoption of these processes (Naudet et al., 2018).

Finally, open code and data are necessary to assess the computational reproducibility of a field. This type of reproducibility increases confidence in the statistical analysis and the integrity of the findings (Eubank, 2016; Hardwicke et al., 2018). Ruling out computational error as a reason for failed translation will allow clinicians to focus on more informative explanations of a trial's results. Unfortunately, the low levels of both data and code sharing would preclude any attempts at assessing the computational reproducibility of the AMOA literature.

Reporting Standards

Masking

Our research revealed that a little over one third of AMOA papers made any mention of masking. This result was significantly higher than that produced by a large-scale survey of the preclinical biomedical literature and, against our expectations, not significantly smaller than the estimate produced by Leung and colleagues (2018) review of anaesthesia, analgesia, and animal welfare (Menke et al., 2020). This indicates that the prevalence of masking in AMOA is good compared to other preclinical areas.

On the other hand, there was no mention of masking in two thirds of articles. This means the findings produced by these experiments may be influenced by bias. The large proportion of experiments that appear not to have implemented masking places the AMOA literature at risk of inflated effect sizes and increased rates of false positives. Given the unequivocal importance of masking in all study designs, why do rates remain low in AMOA research?

A qualitative analysis of attitudes towards masking in preclinical researchers generally is informative (Karp et al., 2022). It revealed a major obstacle was the lack of proficiency in masking techniques, suggesting the need to increase researchers' motivation to engage in the range of educational resources that already exist, such as practical articles and research planning tools (Karanicolas et al., 2010; Munafò et al., 2017; Percie du Sert et al., 2017).

Karp and colleagues (2022) also reported a lack of belief in the value or relevance of masking to the researchers' preclinical area. That these beliefs persist is informative, despite the attention masking has received as part of efforts to improve translation rates (Landis et al., 2012; Moher et al., 2016). Indeed, it is one of the 'Landis 4': four core research aspects that have been targeted for improvement because of the low rates of implementation and the consequences their absence has on translation (Hair et al., 2019; Landis et al., 2012).

Ideally, researchers would use masking because they believe in its value, instead of doing so because of external requirements. Voluntary implementation will likely be of higher quality and it leaves researchers in charge of how research is done (Giner-Sorolla, 2012). However, the low rates of masking in AMOA among other preclinical fields suggest the hoped-for 'cultural change' towards improved transparency and reporting is slow in arriving (Landis et al., 2012; Munafò et al., 2018). This may indicate more involvement on the part of journals and funders in encouraging this change.

The ARRIVE guidelines may present the middle-ground: by requiring statements detailing 'who was aware of the group allocation at different stages', the researcher maintains control over the research process, but the statement allows for greater transparency and thus the possibility of scrutiny about the masking procedure (Percie du Sert, Hurst, et al., 2020). In the current sample, 1.6% of articles included statements of no masking. At the very least, such a statement removes ambiguity. This, of course, is not just relevant to masking but many research design aspects, including all practices measured in this study.

Improving masking is a key focus to improving the 'translational hit' of preclinical research (Landis et al., 2012; Schmidt-Pogoda et al., 2020). Our results demonstrate the field of AMOA has not yet reached acceptable levels of masking and should therefore engage in efforts to improve this practices (Fergusson et al., 2019).

Randomisation

Nearly half of the articles in AMOA reported randomisation. This result was significantly larger than randomisation estimates in Menke and colleagues' (2020) preclinical biomedical literature review and not different from research by Kousholt and colleagues (2022). However, contrary to our expectations, it was lower than estimates from the analgesia, anaesthesia, and animal welfare literature, and the pain and anaesthesia literature (Leung et al., 2018; Fergusson et al., 2019).

While randomisation is the ideal group allocation method and is typically required for most treatments to be ‘proven’, there are times when it may not be appropriate or possible in preclinical research (Bebarta et al., 2003). In these instances, appropriate reporting and defence of such research decisions are required (Bebarta et al., 2003). From this perspective, that nearly 56% of papers included some statement about randomisation or other group allocation method is encouraging.

On the other hand, 44% of AMOA papers made no mention of randomisation or another allocation method.

It is unclear why use of randomisation should not be higher, given that it is not a novel practice and ‘well-established’ randomisation procedures exist (Bespalov et al., 2020; Percie du Sert et al., 2017; Schulz et al., 2016). Bebarta and colleagues (2003) suggested the practice may be considered unnecessary by some animal researchers because of the increased homogeneity in animals compared to humans. This suggests increased efforts are required to highlight the importance of randomisation in reducing bias, balancing confounders, and thus validating the use of inferential statistics (Percie du Sert, Ahluwalia, et al., 2020).

The ARRIVE guidelines 2.0 have attempted to do this with the new ‘Explanation and Elaboration’ section accompanying each reporting requirement (Percie du Sert, Ahluwalia, et al., 2020). Once again, however, journals and funding bodies may need to provide the motivation for some researchers to engage in practices that they have hitherto thought irrelevant.

Randomisation is another of the Landis 4 core reporting requirements (Landis et al., 2012). Despite its importance, the AMOA literature indicates that there is considerable room for improvement in the use of randomisation. As such, widespread attempts to encourage bias minimisation techniques to improve rates of reproducibility and translation are indeed relevant to this field.

Sample size calculation

Most papers in the AMOA literature did not include sample size justifications. 0.8% of papers relied on previous research and 4.9% used a sample size calculation to determine the sample size. This latter estimate was not significantly different to the rates of SSC found in preclinical biomedicine or analgesia, anaesthesia, and animal welfare, although it was significantly smaller than in Kousholt and colleagues' (2018) review of preclinical animal literature (Leung et al., 2018; Menke et al., 2020). Lastly, none of the studies with SSC provided enough detail to recreate the analysis. While in keeping with our expectations, this result is far from optimal.

Evidence shows that several preclinical areas suffer from consistently underpowered studies (Ellis, 2022; Schmidt-Pogoda et al., 2020; Vesterinen et al., 2010). The low prevalence of SSCs found in this study indicates AMOA research may be vulnerable to being underpowered. The lack of reporting of power, however, may preclude a definitive answer (van der Naald et al., 2020).

A potential obstacle for conducting SSCs may be the difficulty in estimating the population parameter with which to carry out the calculation (Flora, 2020). This may be especially true in novel research areas (Schäfer & Schwarz, 2019). While this is legitimate, without reporting this difficulty, the ambiguity about a study's power remains. The large proportion of studies that did not report SSCs means this ambiguity exists in AMOA.

Compulsory reporting about sample size decisions may elucidate this issue, hopefully encouraging researchers to consider robust methods and increasing discussion about the inherent difficulties. Such discourse may also spread awareness of possible solutions, such as the use of the smallest effect size of interest in SSCs (Lakens et al., 2018). Indeed, Nature's mandatory checklist has led to improvements in SSC reporting (M. Macleod, 2019). In instances where practical constraints limit a study's ability to reach appropriate power, multi-laboratory solutions have been suggested(Munafò et al., 2017) (Munafo et al., 2017).

The Landis 4 includes SSC as a design aspect requiring urgent improvement in preclinical research to combat poor translation rates (Landis et al., 2012). Such an improvement would benefit the reproducibility and translation of AMOA.

Data exclusion

A third of papers reported data exclusion and an additional 3% included statements of no exclusion, meaning clarity about included data appeared in nearly 36% of papers. This result was not significantly different to reported data exclusion in preclinical animal research and pain and anaesthesiology, but it was significantly lower than in analgesia and animal welfare research (Fergusson et al., 2019; Kousholt et al., 2022; Leung et al., 2018). Once again, this places AMOA in similar position to other preclinical fields.

Unfortunately, the lack of statements about data exclusion leaves room for ambiguity in 64% of papers. Crucially, research suggests that we cannot assume that data has not been excluded if it is not reported (van der Naald et al., 2020).

Only 16 papers included a statement about outlier exclusion, representing 20% of all reported data exclusion. The inconsistency in defining statistical outliers heightens the need for outlier reporting. Such uncertainty here may contribute to the poor disclosure of this aspect in AMOA research. Increased outlier reporting may have the additional benefit of accelerating progress towards more consistent definitions. Alternatively, working groups of AMOA researchers could generate advice on best practice in defining and handling outliers.

Considering this result together with the low rates of SSCs raises the concern that the consequences of unreported data exclusion may be exacerbated by the potential for AMOA to be underpowered in line with other preclinical fields (Schmidt-Pogoda et al., 2020; Vesterinen et al., 2010). Further, the limited sharing of data and code in AMOA would make unreported data exclusion hard to detect. Lastly, given the lack of preregistration, it is unclear if researchers are protecting against the potential for biased data removal by deciding on *a priori* exclusion criteria and handling procedures.

Data handling is the final aspect included in the Landis 4. Poor reporting of exclusions has negative consequences for the methods and results reproducibility of AMOA research, and the predictive value for later translation (Landis et al., 2012).

Landis 4

The reporting of all four aspects of rigorous research is disappointing. Only four out of 247 articles made some mention of randomisation, masking, SSC, and data exclusion – including statements of non-implementation. Despite these issues being definitively elucidated more than a decade ago, the continued absence of these measures in much of AMOA research may diminish translation potential (Landis et al., 2012).

Multiple comparisons adjustments

Our study found that three quarters of papers reported a MCA. This result is significantly larger than estimates from cardiovascular and analgesic clinical trials (Gewandter et al., 2014; Khan et al., 2020). This is an encouraging result, showing that AMOA is doing substantially better than existing estimates.

MCA is particularly important in research that may lead to treatment development or influence policy, as there is a greater cost of discovering a false positive compared to a false negative (Althouse, 2016). These considerations are relevant to AMOA given the implications of prescription opioid addiction for government and health administration.

Despite our best efforts to estimate the upper bound the prevalence of all target measures, it is likely that we were unable to capture all instances of MCA because of the variety of types. This means the true estimate may be even higher.

Closing the remaining gap to perfect reporting may require only minimal encouragement from reviewers and journals given the value of MCA is clearly appreciated. Indeed, it appears the concern that multiplicity is being ‘widely ignored’ in psychology is not relevant to AMOA (Cramer et al., 2016).

ARRIVE compliance

Only 7.3% of articles reported compliance to the ARRIVE guidelines. This may be unsurprising considering that only seven of the top 25 journals in our sample endorsed the ARRIVE guidelines (ARRIVE Guidelines, 2023).

Moreover, of the 18 articles that stated compliance with ARRIVE, only one reported all items in the Essential 10 recommendations. We do not suggest this is an exhaustive evaluation of the relationship between stated compliance and actual compliance, however it does raise the question whether purported compliance with ARRIVE improves reporting standards. There is mixed evidence about whether journal endorsement leads to better reporting (Baker et al., 2014; Hair et al., 2019; Hepkema et al., 2022). Conversely, Nature's reporting checklist that is followed up by reviewers led to improvements in Landis 4 reporting (Han et al., 2017; M. Macleod, 2019). This research suggests greater involvement of journals and reviewers during the prepublication process is required. While this may seem out of reach given reviewers are often already 'overextended', Landis and colleagues (2012) suggest that clear requirements of a manuscript will make reviewers' jobs easier.

Accurate Reporting

Test statistic accuracy

We were able to detect test statistics in 75% of articles which was a significantly larger proportion than that found by Nuijten and colleagues (2016). This means the formatting of test statistics was consistently in line with APA formatting, facilitating efficient accuracy checks.

Of the papers where test statistics were detected, 52% had at least one non-decision error. Moreover, 11% of papers contained one or more decision errors. Neither of these results were significantly different to the comparable results found in psychology (M. B. Nuijten et al., 2016). The low rates of data sharing in AMOA would make correcting these inaccuracies difficult (Nuijten et al., 2016)

Given the importance of statistical significance in evaluating preclinical research for further investigation, and the ease with which accuracy can be verified, these rates may be unacceptable.

Some of these inaccuracies may be deliberate, incentivised by publication bias (John et al., 2012; Nuijten et al., 2016). On the other hand, researchers are vulnerable to typographical and other basic errors that can cause such inconsistencies (Hardwicke et al., 2018). Either alternative is good motivation to introduce pre-publication reporting accuracy checks as a matter of course. Given the large proportion of papers where test statistics were detected, AMOA is in a good position to pioneer such a self-correcting practice (Vazire & Holcombe, 2022)

statcheck, however, is not a perfect tool and is likely to have missed test statistics even in papers where some were detected. We cannot be sure how these missed test statistics would influence our results. These limitations, however, may be secondary to the principal point: errors persist in the AMOA literature and, given the availability of tools to rapidly verify test statistics, this should not be the case.

Our results demonstrate that there remains considerable room for improvement in the AMOA literature. The low rates of transparency measures reflect the slow uptake of these practices designed to combat biased and irreproducible research. These practices may benefit the robustness of the AMOA findings, which is likely needed given the currently low prevalence of reporting of bias minimisation practices and the persistence of test statistic inconsistencies. This is cause for concern, as similar findings in preclinical areas have been associated with inflated effect sizes and a higher preponderance of false positives (Sena et al., 2010; Vesterinen et al., 2010). These are detrimental to the reproducibility or translation potential of a literature.

Solutions

There appears to be tension in proposed solutions to the issues examined between enforcing change and waiting for voluntary change. There may be a middle ground, however, in enforcing reporting of implementation or non-implementation of crucial practices, as recommended by ARRIVE. This solution enjoys both benefits of researcher freedom and allowing for informed assessment of research. Encouraging awareness in this way may accelerate voluntary change (Munafò et al., 2018).

While such a remedy seems simple, attempts at implementing the ARRIVE guidelines have proved otherwise (Hair et al., 2019; Hepkema et al., 2022; Ting et al., 2015). As such, the involvement of all stakeholders in AMOA is imperative. Indeed, any solution that rests solely on the researchers and reviewers will not be sustainable.

Novel solutions may arise from further research into researcher attitudes and perceived barriers to implementing new and not-so-new. Targeting such research at the level of individual fields may be particularly productive.

Limitations and strengths

Several limitations to this study should be considered. Firstly, when coding an article, we did not ascertain whether the target characteristics were appropriate to the design, nor whether they were applied in all the instances required. This was largely due to constraints in coder ability. However, the target aspects were selected for their widespread applicability in hypothesis-testing research. Further, statements of non-implementation are recommended for the characteristics studied here to clarify this very issue (Percie du Sert, Hurst, et al., 2020).

A second limitation arose because of our decision to review the online versions of articles. We hoped this would enable better detection of hyperlinks and supplementary files. However, because articles were often available through several databases, it was not ensured that coders were accessing the original version of the article. This led to the unexpected obstacle of finding conflicting results for a given article. To reduce the impact of this, coder 1 checked all available article versions when resolving coder discrepancies. We cannot rule out the possibility, though, that there were some instances where both original coders missed a target characteristic, for this reason or simple through human error.

Similarly, our search terms likely did not capture all the possible permutations of a target characteristic.

These limitations reflect threats to the accuracy of our estimates. To compensate to some extent for these insufficiencies, we adopted a charitable coding stance so that our approach would not be seen to unfairly criticise the AMOA literature.

Importantly, these limitations do not detract from the principal goal of this study: to determine whether attempts to improve transparency and reproducibility measures are relevant to AMOA. We believe we succeeded in this goal.

A strength of this study is that captured all articles returned by our search. This will improve the accuracy of our estimates as we do not have to consider the effects of sampling variability. This contrasts with similar metaresearch that often randomly samples from a larger pool. As such, our results may be somewhat useful for generalising to fields with similar methodological approaches.

In another departure from conventional metaresearch, we addressed a relatively small research field. We hope that by doing so, our results are more directly applicable and thereby actionable.

Generalisability

Despite our tentative optimism about the generalisability of these findings, we believe these results would better serve as motivation for related fields to conduct reviews of their own literature. This is imperative as each field has distinct factors to consider, despite methodological or theoretical similarities.

Implications and conclusion

The implications of our results for the reproducibility of the AMOA literature are cause for concern. Attempts at computational reproductions would be precluded by the low rates of access to the original data and code. Methods reproducibility is at least partly obscured by the poor reporting of randomisation, masking, data handling, and adjustments for multiple comparisons. Finally, the lack of widespread reporting of bias minimisation practices, combined with unclear power places this field at risk of poor results reproducibility (Open Science Collaboration, 2015). These results may affect an effect’s translation potential, and impede clinicians from being able to count on robust preclinical methodology (Landis et al., 2012).

Further, despite the possibility that AMOA is considered a ‘harder’ psychological field, it remains in danger of poor reproducibility. This means efforts taken in other fields of psychology in response to the replication crisis are indeed relevant – and needed – in AMOA research.

Moving forward, it is ideal for AMOA researchers to lead the charge on improving transparency and reporting standards in their own field. This would allow those that best understand the nuances of the field to shape it.

This study contributes the first metascience study in animal models of addiction. By focusing on opioid research alone, we hope to spur change by contributing findings that are meaningful and immediately applicable to researchers working in this area.

References

Adewumi, M. T., Vo, N., Tritz, D., Beaman, J., & Vassar, M. (2021). An evaluation of the practice of transparency and reproducibility in addiction medicine literature. *Addictive Behaviors*, *112*, Article 106560. https://doi.org/10.1016/j.addbeh.2020.106560

Althouse, A. D. P. (2016). Adjust for multiple comparisons? It's not that simple. *The Annals of Thoracic Surgery*, *101*(5), 1644–1645. https://doi.org/10.1016/j.athoracsur.2015.11.024

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, *28*(1), 5–21. https://doi.org/10.1016/j.leaqua.2017.01.006

ARRIVE Guidelines. (2023). *Journals*. https://arriveguidelines.org/supporters/journals

Australian Institute of Health, & Welfare. (2018). *Opioid harm in Australia: And comparisons between Australia and Canada*. AIHW.

Baker, D., Lidster, K., Sottomayor, A., & Amor, S. (2014). Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLOS Biology*, *12*(1), Article e1001756. https://doi.org/10.1371/journal.pbio.1001756

Bebarta, V., Luyten, D., & Heard, K. (2003). Emergency medicine animal research: Does use of randomization and blinding affect the results? *Academic Emergency Medicine*, *10*(6), 684–687. https://doi.org/10.1197/aemj.10.6.684

Bergkvist, L. (2020). Preregistration as a way to limit questionable research practice in advertising research. *International Journal of Advertising*, *39*(7), 1172–1180. https://doi.org/10.1080/02650487.2020.1753441

Bespalov, A., Michel, M. C., & Steckler, T. (2020). Blinding and Randomization. *Handbook of Experimental Pharmacology*, *257*, 81–100. https://doi.org/10.1007/164_2019_279

Best, L., Smith, L., & Stubbs, D. (2001). Graph use in psychology and other sciences. *Behavioural Processes*, *3*(54), 155–156. https://doi.org/10.1016/s0376-6357(01)00156-5.

Bonett, D. G., & Price, R. M. (2012). Adjusted wald confidence interval for a difference of binomial proportions based on paired data. *Journal of Educational and Behavioral Statistics*, *37*(4), 479–488. https://doi.org/10.3102/1076998611411915

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Centre for Open Science. (2023). *Registered reports: Peer review before results are known to align scientific values and practices.* https://www.cos.io/initiatives/registered-reports

Chin, J. (2023). The transparency of quantitative empirical legal research published in highly ranked law journals (2018–2020): An observational study. *F1000 Research*, *12*(144), 23. https://doi.org/10.12688/f1000research.127563.1

Cook, C. N., Freeman, A. R., Liao, J. C., & Mangiamele, L. A. (2022). The philosophy of outliers: Reintegrating rare events into biological science. *Integrative and Comparative Biology*, *61*(6), 2191–2198. https://doi.org/10.1093/icb/icab166

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L. J., & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

Crossley, N. A., Sena, E., Goehler, J., Horn, J., Van Der Worp, B., Bath, P. M. W., MacLeod, M., & Dirnagl, U. (2008). Empirical evidence of bias in the design of experimental stroke studies: A metaepidemiologic approach. *Stroke*, *39*(3), 929–934. https://doi.org/10.1161/STROKECoder 4A.107.498725

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), Article 200805. https://doi.org/10.1098/rsos.200805

Ellis, R. J. (2022). Questionable research practices, low statistical power, and other obstacles to replicability: Why preclinical neuroscience research would benefit from registered reports. *eNeuro*, *9*(4), Article ENEURO.0017-22.2022. https://doi.org/10.1523/ENEURO.0017-22.2022

Epskamp, S., & Nuijten, M. B. (2015). statcheck: Extract statistics from articles and recompute p values. *R Package Version 1.0.1.* http:// CRAN.R-project.org/package=statcheck

Epstein, D. H., Heilig, M., & Shaham, Y. (2018). Science-based actions can help address the opioid crisis. *Trends in Pharmacological Sciences (Regular Ed.)*, *39*(11), 911–916. https://doi.org/10.1016/j.tips.2018.06.002

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*. https://doi.org/10.7554/eLife.67995

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*. https://doi.org/10.7554/eLife.71601

Eubank, N. (2016). Lessons from a decade of replications at the quarterly journal of political science. *Political Science & Politics*, *49*(2), 273–276. https://doi.org/10.1017/S1049096516000196

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PloS One*, *5*(4), Article e10068. https://doi.org/10.1371/journal.pone.0010068

Fergusson, D. A., Avey, M. T., Barron, C. C., Bocock, M., Biefer, K. E., Boet, S., Bourque, S. L., Conic, I., Chen, K., Dong, Y. Y., Fox, G. M., George, R. B., Goldenberg, N. M., Gragasin, F. S., Harsha, P., Hong, P. J., James, T. E., Larrigan, S. M., MacNeil, J. L., … Lalu, M. M. (2019). Reporting preclinical anesthesia study (REPEAT): Evaluating the quality of reporting in the preclinical anesthesiology literature. *PloS One*, *14*(5), Article e0215221. https://doi.org/10.1371/journal.pone.0215221

Flora, D. B. (2020). Thinking about effect sizes: From the replication crisis to a cumulative psychological science. *Canadian Psychology*, *61*(4), 318–330. https://doi.org/10.1037/cap0000218

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), Article e109019. https://doi.org/10.1371/journal.pone.0109019

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*, 1–17. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gewandter, J. S., Smith, S. M., McKeown, A., Burke, L. B., Hertz, S. H., Hunsinger, M., Katz, N. P., Lin, A. H., McDermott, M. P., Rappaport, B. A., Williams, M. R., Turk, D. C., &

Dworkin, R. H. (2014). Reporting of primary analyses and multiplicity adjustment in recent analgesic clinical trials: ACTTION systematic review and recommendations. *Pain (Amsterdam)*, *155*(3), 461–466. https://doi.org/10.1016/j.pain.2013.11.009

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Gomes, D. G. E., Pottier, P., Crystal-Ornelas, R., Hudgins, E. J., Foroughirad, V., Sánchez-Reyes, L. L., Turba, R., Martinez, P. A., Moreau, D., Bertram, M. G., Smout, C. A., & Gaynor, K. M. (2022). Why don’t we share data and code? Perceived barriers and benefits to public archiving practices. *Proceedings of the Royal Society. B, Biological Sciences*, *289*(1987), Article 20221113. https://doi.org/10.1098/rspb.2022.1113

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Getting to Good: Research Integrity in the Biomedical Sciences*, *8*(341), 96–102. https://doi.org/10.1126/scitranslmed.aaf5027

Gorman, D. M. (2019). Use of publication procedures to improve research integrity by addiction journals. *Addiction (Abingdon, England)*, *114*(8), 1478–1486. https://doi.org/10.1111/add.14604

Green, C. D., Abbas, S., Belliveau, A., Beribisky, N., Davidson, I. J., DiGiovanni, J., Heidari, C., Martin, S. M., Oosenbrug, E., & Wainewright, L. M. (2018). Statcheck in Canada: What proportion of CPA Journal articles contain errors in the reporting of *p*-values? *Canadian Psychology = Psychologie Canadienne*, *59*(3), 203–210. https://doi.org/10.1037/cap0000139

Hair, K., Macleod, M. R., Wever, K. E., Tanriver-Ayder, E., & Sena, E. S. (2019). A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines

(IICARus). *Research Integrity and Peer Review*, *4*(1), 12–12. https://doi.org/10.1186/s41073-019-0069-3

Hamilton, D. G., Hong, K., Fraser, H., Rowhani-Farid, A., Fidler, F., & Page, M. J. (2023). Prevalence and predictors of data and code sharing in the medical and health sciences: Systematic review with meta-analysis of individual participant data. *BMJ*, *382*, Article e075767. https://doi.org/10.1136/bmj-2023-075767

Han, S., Olonisakin, T. F., Pribis, J. P., Zupetic, J., Yoon, J. H., Holleran, K. M., Jeong, K., Shaikh, N., Rubio, D. M., & Lee, J. S. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PloS One*, *12*(9), Article e0183591. https://doi.org/10.1371/journal.pone.0183591

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PloS One*, *13*(8), Article e0201856. https://doi.org/10.1371/journal.pone.0201856

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), Article 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). *Royal Society Open Science*, *7*(2), Article 190806. https://doi.org/10.1098/rsos.190806

Heirene, R. M. (2021). A call for replications of addiction research: Which studies should we replicate and what constitutes a “successful” replication? *Addiction Research & Theory*, *29*(2), 89–97. https://doi.org/10.1080/16066359.2020.1751130

Hepkema, W. M., Horbach, S. P. J. M., Hoek, J., & Halffman, W. (2022). Misidentified biomedical resources: Journal guidelines are not a quick fix. *International Journal of Cancer*, *150*(8), 1233–1243. https://doi.org/10.1002/ijc.33882

Hirst, J. A., Howick, J., Aronson, J. K., Roberts, N., Perera, R., Koshiaris, C., & Heneghan, C. (2014). The need for randomization in animal trials: An overview of systematic reviews. *PloS One*, *9*(6), Article e98856. https://doi.org/10.1371/journal.pone.0098856

Holman, C., Piper, S. K., Grittner, U., Diamantaras, A. A., Kimmelman, J., Siegerink, B., & Dirnagl, U. (2016). Where have all the rodents gone? The effects of attrition in experimental research on cancer and stroke. *PLOS Biology*, *14*(1), Article e1002331. https://doi.org/10.1371/journal.pbio.1002331

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*, *344*(7848), 20–20. https://doi.org/10.1136/bmj.e1119

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 8. https://doi.org/10.1177/0956797611430953

Karanicolas, P. J., Farrokhyar, F., & Bhandari, M. (2010). Practical tips for surgical research: Blinding: Who, what, when, why, how? *Canadian Journal of Surgery. Journal Canadien de Chirurgie*, *53*(5), 345–348.

Karp, N. A., Pearl, E. J., Stringer, E. J., Barkus, C., Ulrichsen, J. C., & Percie du Sert, N. (2022). A qualitative study of the barriers to using blinding in in vivo experiments and suggestions for improvement. *PLOS Biology*, *20*(11), Article e3001873. https://doi.org/10.1371/journal.pbio.3001873

Kerr, N. L. (1998). HARKing: Hypothesising After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 21. https://doi.org/10.1207/s15327957pspr0203_4

Khan, M. S., Khan, M. S., Ansari, Z. N., Siddiqi, T. J., Khan, S. U., Riaz, I. B., Asad, Z. U. A., Mandrola, J., Wason, J., Warraich, H. J., Stone, G. W., Bhatt, D. L., Kapadia, S. R., & Kalra, A. (2020). Prevalence of multiplicity and appropriate adjustments among cardiovascular randomized clinical trials published in major medical journals. *JAMA Netw Open*, *3*(4), Article e203082. https://doi.org/10.1001/jamanetworkopen.2020.3082

Kimmelman, J., & Anderson, J. A. (2012). Should preclinical studies be registered? *Nature Biotechnology*, *30*(6), 488–489. https://doi.org/10.1038/nbt.2261

Kousholt, B. S., Præstegaard, K. F., Stone, J. C., Thomsen, A. F., Johansen, T. T., Ritskes-Hoitinga, M., & Wegener, G. (2022). Reporting quality in preclinical animal experimental research in 2009 and 2018: A nationwide systematic investigation. *PloS One*, *17*(11), e0275962–e0275962. https://doi.org/10.1371/journal.pone.0275962

Kubina, R. M., Kostewicz, D. E., & Datchuk, S. M. (2008). An initial survey of fractional graph and table area in behavioral journals. *Behaviour Analyst*, *31*(1), 61–66. https://doi.org/10.1007/bf03392161

Lakens, D., Scheel, A. M., & Isager, P. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G., Darnell, R. B., Ferrante, R. J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H. E., Golub, R. M., Goudreau, J. L., Gross, R. A., Gubitz, A. K., Hesterlee, S. E., Howells, D. W., … Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature (London)*, *490*(7419), 187–191. https://doi.org/10.1038/nature11556

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, Article 104332. https://doi.org/10.1016/j.jml.2022.104332

Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is “N” in cell culture and animal experiments? *PLOS Biology*, *16*(4) Article e2005282. https://doi.org/10.1371/journal.pbio.2005282

Leek, J. T., & Storey, J. D. (2008). General framework for multiple testing dependence. *Proceedings of the National Academy of Sciences - PNAS*, *105*(48), 18718–18723. https://doi.org/10.1073/pnas.0808709105

Leung, V., Rousseau-Blass, F., Beauchamp, G., & Pang, D. S. J. (2018). Arrive has not arrived: Support for the arrive (animal research: Reporting of in vivo experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PloS One*, *13*(5), Article e0197882. https://doi.org/10.1371/journal.pone.0197882

Lowenstein, P. R., & Castro, M. G. (2009). Uncertainty in the translation of preclinical experiments to clinical trials. Why do most phase III clinical trials fail? *Current Gene Therapy*, *9*(5), 368–374. https://doi.org/10.2174/156652309789753392

Macleod, M. (2019). Did a change in Nature journals' editorial policy for life sciences research improve reporting? *BMJ Open Science*, *3*(1), Article e000035. https://doi.org/10.1136/bmjos-2017-000035

Macleod, M. R., Lawson McLean, A., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., Hirst, T., Hemblade, R., Bahor, Z., Nunes-Fonseca, C., Potluru, A., Thomson, A., Baginskitae, J., Egan, K., Vesterinen, H., Currie, G. L., Churilov, L., Howells, D. W., & Sena, E. S. (2015). Risk of bias in reports of in vivo research: A focus for improvement. *PLOS Biology*, *13*(10), 1–12. https://doi.org/10.1371/journal.pbio.1002273

Macleod, M. R., van der Worp, H. B., Sena, E. S., Howells, D. W., Dirnagl, U., & Donnan, G. A. (2008). Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke (1970)*, *39*(10), 2824–2829. https://doi.org/10.1161/STROKECoder 4A.108.515957

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: how often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Menke, J., Roelandse, M., Ozyurt, B., Martone, M., & Bandrowski, A. (2020). The rigor and transparency index quality metric for assessing biological and medical science methods. *iScience*, *23*(11), Article 101698. https://doi.org/10.1016/j.isci.2020.101698

Miller, J. (2023). Outlier Exclusion Procedures for Reaction Time Analysis: The Cures Are Generally Worse Than the Disease. *Journal of Experimental Psychology: General*. https://doi.org/doi.org/10.1037/xge0001450

Moher, D. D., Glasziou, P. F., Chalmers, I. Ds., Nasser, M. D. D. S., Bossuyt, P. M. M. P., Korevaar, D. A. M. D., Graham, I. D. P., Ravaud, P. P., & Boutron, I. P. (2016). Increasing value and reducing waste in biomedical research: Who's listening? *The Lancet (British Edition)*, *387*(10027), 1573–1586. https://doi.org/10.1016/S0140-6736(15)00307-4

Munafò, M. R. (2015). Opening up addiction science. *Society for the Study of Addiction*, *111*, 387–388. https://doi.org/10.1111/add.13147

Munafò, M. R., Hollands, G. J., & Marteau, T. M. (2018). Open science prevents mindless science. *BMJ (Online)*, *363*, Article k4309. https://doi.org/10.1136/bmj.k4309

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. https://doi.org/10.1038/s41562-016-0021

National Institutes of Health. (2023a). *Discovery of novel targets for pain treatment*. https://heal.nih.gov/research/preclinical-translational/novel-targets

National Institutes of Health. (2023b). *Drug overdose death rates*. https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in The BMJ and PLOS Medicine. *BMJ*, *360*, 1–11. https://doi.org/10.1136/bmj.k400

Niso, G., Krol, L. R., Combrisson, E., Dubarry, A. S., Elliott, M. A., François, C., Héjja-Brichard, Y., Herbst, S. K., Jerbi, K., Kovic, V., Lehongre, K., Luck, S. J., Mercier, M., Mosher, J. C., Pavlov, Y. G., Puce, A., Schettino, A., Schön, D., Sinnott-Armstrong, W., … Chaumon, M. (2022). Good scientific practice in EEG and MEG research: Progress and perspectives. *NeuroImage*, *257*, Article 119056. https://doi.org/10.1016/j.neuroimage.2022.119056

Norris, E., He, Y., Loh, R., West, R., & Michie, S. (2021). Assessing markers of reproducibility and transparency in smoking behaviour change intervention evaluations. *Journal of Smoking Cessation*, *2021*, 1–12. https://doi.org/10.1155/2021/6694386

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences - PNAS*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Nuijten, M., & Polanin, J. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, *11*(5), 574–579. https://doi.org/10.1002/jrsm.1408

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (American Association for the Advancement of Science)*, *349*(6251), 943–943.

Owzar, K., Barry, W. T., & Jung, S.-H. (2011). Statistical considerations for analysis of microarray experiments. *Clinical and Translational Science*, *4*(6), 466–477. https://doi.org/10.1111/j.1752-8062.2011.00309.x

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pennington, C. R. (2023). Open data through Registered Reports can accelerate cumulative knowledge. *Addiction Research & Theory*, *31*(3), 155–156. https://doi.org/10.1080/16066359.2023.2176848

Percie du Sert, N., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Hurst, V., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Würbel, H. (2020). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLOS Biology*, *18*(7), Article e3000411. https://doi.org/10.1371/journal.pbio.3000411

Percie du Sert, N., Bamsey, I., Bate, S. T., Berdoy, M., Clark, R. A., Cuthill, I., Fry, D., Karp, N. A., Macleod, M., Moon, L., Stanford, S. C., & Lings, B. (2017). The Experimental Design Assistant. *Nature Methods*, *15*(9), Article 1024-e2003779. https://doi.org/10.1371/journal.pbio.2003779

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Boutron, I. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology*, *18*(7), 12. https://doi.org/10.1371/journal.pbio.3000410

Perrin, S. (2014). Preclinical research: Make mouse studies work. *Nature (London)*, *507*(7493), 423–425. https://doi.org/10.1038/507423a

Pui Yu Lee. (2022). *Estimating the Prevalence of Direct Replication Articles and Journal Policies in Psychology: 2010-2021* [Unpublished honour's thesis].

Quentin André. (2023). *Outlier Exclusion Procedures Must be Blind to the Researcher's Hypothesis* [Manuscript]. https://osf.io/3tz76/

Riley, S. P., Swanson, B., Brismée, J.-M., & Sawyer, S. F. (2016). A systematic review of orthopaedic manual therapy randomized clinical trials quality. *The Journal of Manual & Manipulative Therapy*, *24*(5), 241–252. https://doi.org/10.1080/10669817.2015.1119372

Rooke, E. D. M., Vesterinen, H. M., Sena, E. S., Egan, K. J., & Macleod, M. R. (2011). Dopamine agonists in animal models of Parkinson's disease: A systematic review and meta-analysis. *Parkinsonism & Related Disorders*, *17*(5), 313–320. https://doi.org/10.1016/j.parkreldis.2011.02.010

Rubin, M. (2017). Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, *21*(3), 269–275. https://doi.org/10.1037/gpr0000123

Sauro, J. (2023). *Confidence interval calculator for a completion rate*. https://measuringu.com/calculators/wald/

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007467

Schmidt-Pogoda, A., Bonberg, N., Koecke, M. H. M., Strecker, J.-K., Wellmann, J., Bruckmann, N.-M., Beuker, C., Schäbitz, W.-R., Meuth, S. G., Wiendl, H., Minnerup, H., & Minnerup, J. (2020). Why most acute stroke studies are positive in animals but not in patients: A systematic comparison of preclinical, early phase, and phase 3 clinical trials of neuroprotective agents. *Annals of Neurology*, *87*(1), 40–51. https://doi.org/10.1002/ana.25643

Schulz, J. B., Cookson, M. R., & Hausmann, L. (2016). The impact of fraudulent and irreproducible data to the translational research crisis: Solutions and implementation. *Journal of Neurochemistry*, *139*(S2), 253–270. https://doi.org/10.1111/jnc.13844

Sena, E. S., Bart van der Worp, H., Bath, P. M. W., Howells, D. W., & Macleod, M. R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLOS Biology*, *8*(3), Article e1000344. https://doi.org/10.1371/journal.pbio.1000344

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, Y. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 14. https://doi.org/10.1037/a0033242

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social Studies of Science*, *30*(1), 73–94. https://doi.org/10.1177/030631200030001003

Statskingdom. (2022a, June). *Statistics online*. https://www.statskingdom.com/index.html

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PloS One*, *8*(6), Article e67111. https://doi.org/10.1371/journal.pone.0067111

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), 18. https://doi.org/10.1371/journal.pbio.2000797

Talylor & Francis. (2023). *Author services: Supporting Taylor & Francis authors*. https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/data-availability-statements/

Tikka, C., Verbeek, J., Ijaz, S., Hoving, J. L., Boschman, J., Hulshof, C., & de Boer, A. G. (2021). Quality of reporting and risk of bias: A review of randomised trials in occupational health. *Occupational and Environmental Medicine (London, England)*, *78*(9), 691–696. https://doi.org/10.1136/oemed-2020-107038

Ting, K. H. J., Hill, C. L., & Whittle, S. L. (2015). Quality of reporting of interventional animal studies in rheumatology: A systematic review using the ARRIVE guidelines. *International Journal of Rheumatic Diseases*, *18*(5), 488–494. https://doi.org/10.1111/1756-185X.12699

Uher, J. (2021). Psychology’s status as a science: Peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integrative Physiological and Behavioral Science*, *55*(1), 212–224. https://doi.org/10.1007/s12124-020-09545-0

van der Naald, M., Chamuleau, S. A. J., Menon, J. M. L., de Leeuw, W., de Haan, J. J., Duncker, D. J., & Wever, K. E. (2021). A 3-year evaluation of preclinicaltrials.eu reveals room for improvement in preregistration of animal studies. *PLOS Biology*, *19*(9), Article e3001397. https://doi.org/10.1371/journal.pbio.3001397

van der Naald, M., Wenker, S., Doevendans, P. A., Wever, K. E., & Chamuleau, S. A. J. (2020). Publication rate in preclinical research: A plea for preregistration. *BMJ Open Science*, *4*(1), Article e100051. https://doi.org/10.1136/bmjos-2019-100051

Vassar, M., Jellison, S., Wendelbo, H., & Wayant, C. (2020). Data sharing practices in randomized trials of addiction interventions. *Addictive Behaviors*, *102*, Article 106193. https://doi.org/10.1016/j.addbeh.2019.106193

Vassar, M., Roberts, W., Cooper, C. M., Wayant, C., & Bibens, M. (2020). Evaluation of selective outcome reporting and trial registration practices among addiction clinical trials. *Addiction (Abingdon, England)*, *115*(6), 1172–1179. https://doi.org/10.1111/add.14902

Vazire, S., & Holcombe, A. O. (2022). Where are the self-correcting mechanisms in science? *Review of General Psychology*, *26*(2), 212–223. https://doi.org/10.1177/10892680211033912

Venniro, M., Banks, M. L., Heilig, M., Epstein, D. H., & Shaham, Y. (2020). Improving translation of animal models of addiction and relapse by reverse translation. *Nature Reviews. Neuroscience*, *21*(11), 625–643. https://doi.org/10.1038/s41583-020-0378-z

Vesterinen, H. M., Sena, E. S., ffrench-Constant, C., Williams, A., Chandran, S., & Macleod, M. R. (2010). Improving the translational hit of experimental treatments in multiple sclerosis. *Multiple Sclerosis*, *16*(9), 1044–1055. https://doi.org/10.1177/1352458510379612

Watzlawick, R., Antonic, A., Sena, E. S., Kopp, M. A., Rind, J., Dirnagl, U., Macleod, M., Howells, D. W., & Schwab, J. M. (2019). Outcome heterogeneity and bias in acute experimental spinal cord injury: A meta-analysis. *Neurology*, *93*(1), e40–e51. https://doi.org/10.1212/WNL.0000000000007718

WebPlotDigitizer. (2022). *Web based tool to extract data from plots, images, and maps*. https://automeris.io/WebPlotDigitizer/

Appendix A

Original preregistration

Title

Smelling a rat: low rates of open science and anti-bias practices in animal models of opioid addiction/Can we trust this research?

Study rationale & aim

Scrutiny of research practices is an essential part of a credible, self-correcting science. Scrutiny is facilitated by transparency and detailed reporting in the research process, which allows for reproducible, and therefore robust, results. Despite the benefits of open research practices, rates are low across many fields, including psychology and neuroscience. This has contributed to the 'replication crisis'. Steps to address the underlying issues have been adopted to some extent, but currently, we know little about the state of the field in many specific subdisciplines. One such subdiscipline is animal models of opioid addiction. The current study will estimate the prevalence of open science practices in this research literature, including rates of preregistration, registered reports, compliance with the ARRIVE guidelines, replication studies, and availability of raw data and analysis scripts. Further, the plan is for levels of masking, randomisation and outlier exclusion to be collected, and use of power analyses to calculate sample sizes and multiple comparison corrections in data analysis. Lastly, the *p*-values associated with statistical tests in each paper are counted and checked for inconsistency with the reported test statistic.

Methods: Qualitative Study

To get an indication of average effect sizes in the animal models of opioid addiction literature, 14 meta-analyses and systematic reviews will be analysed. Aside from the effect sizes, we will also look at any bias assessments carried out in the papers.

Hypotheses. This study is observational and exploratory. As such we do not have hypotheses. We do, however, expect the effect sizes to be moderate to large.

Search string used to generate sample. addict* OR substance abuse OR drug addiction OR drug treatment AND opioid OR opiate OR heroin AND behaviour* OR behavior*

Search procedure. We searched Scopus, Web of Science, PSYCinfo and PubMed. We limited results to "meta-analysis" or "systematic review", to "animal" and written in English. Results were also limited to being published between 2013 and 2023. This search returned 8 results on Scopus, 11 from Web of Science, 12 on PSYCinfo and 2 from PubMed. After removing duplicates, 29 journal articles remained. Initial screening (reading the abstracts) excluded 15: 2 were not systematic reviews or meta-analyses; 1 was not relevant to opioid use; 12 did not include preclinical research. That left a total of 14 papers. This sample was finalised on the 25th July, 2023.

Sample size rationale. We were unsure how many reviews our search would return. We were prepared to take a random sample of about 15-20 meta-analyses/systematic reviews from a larger pool. However, as the search has returned 14 reviews, we are able to analyse all search results.

Pilot coding. The first round of pilot coding by Coder 2 looked at meta-analyses from clinical and preclinical research. From this process it was understood that effect sizes may need to be converted so that they are comparable across meta-analyses/systematic reviews.

A further round of pilot coding by Coder 2 revealed a variety of methods are used to assess different types of bias. As such, the codebook designed for the qualitative study is unrestrictive. We do not expect to be able to extract the same characteristics from each review.

Data extraction. The qualitative nature of this study means these papers will be read and relevant information will be extracted and categorised as relating to effect sizes or bias estimation. Other relevant details may be noted, such as percentage of papers reporting randomisation. There are likely to be missing values, but due to the qualitative nature of this study, we do no foresee this to be a problem.

Inclusion criteria: published between January 2013 and August 2023; reviews studies related to opioid use (including pain research); studies included in the reviews used behavioural paradigms**;** reviews preclinical literature (can also include clinical literature)**;** *in vivo* research

Data analysis. We will convert effect sizes into a single standardised effect size type. We will choose the standardised effect size type according to which is most commonly used in the sample. Pilot coding indicates this is likely to be Hedge's *g*.

Summary of what data has been collected or looked at prior to posting the preregistration:

Study 1: Qualitative: sample has been located. N = 14. Data extraction has not begun.

Study 2: Quantitative: sample has been located. N = 262. Data extraction has begun. Despite the original search being carried out on the 12th of June, coding remains in the early stages: as of the 8th August 156 papers have been coded by Coder 1; 15 by Coder 4; 22 by Coder 3; 12 by Coder 2. Note that the papers are coded in duplicate. The plan is for Coder 1 to code the entire sample and the other three coders to each code a third.

Methods: Quantitative Study

To estimate the prevalence of thorough reporting and open science practices in the animal models of opioid addiction (AMOA) literature, we will examine journal articles published between 2019 and 2023.

Hypotheses. This study is observational in nature and therefore we do not have hypotheses per se. In keeping with the preliminary research addressing this question in addiction research, however, we expect rates of open science practices and compliance with reporting guidelines ARRIVE to be low (Adewumi et al., 2021).

Search string used to generate sample. addict* OR substance abuse OR drug addiction OR drug treatment AND opioid OR opiate OR heroin AND treatment OR treat* AND behaviour* OR behavior*

Search process. We searched Scopus, Web of Science, PSYCinfo and PubMed. We limited results to "article" or "empirical study", to "animal", written in English and published between 2019 and 2023. Search results were then limited to research areas in Scopus ("neuroscience", "psychology" and "multidisciplinary"). The other databases either didn't have subject areas to choose from after the search had been run (PSYCinfo, PubMed), or all subject areas suggested seemed relevant (Web of Science).

This search returned 123 results on Scopus, 64 from Web of Science, 53 on PSYCinfo and 125 from PubMed. After removing duplicates, 262 journal articles remain.

Upon preparing this preregistration it was noticed that the subject area "pharmacology, toxicology and pharmaceutics" was not included in the Scopus database subject areas. Including this subject area added an additional 82 papers once duplicates had been removed. This was an oversight. To rectify this, instead of including studies that looked at other substance use disorders but still appeared in our results (because of investigation into an opioid receptor agonist, for example) we decided to exclude these. We will replace them with studies appearing under the "pharmacology, toxicology and pharmaceutics" subject area. We aim to include as many as possible in the time constraints of the coders.

Alerts were set up on all databases except for PubMed to notify of relevant papers published during coding time (June-August). Attempts to set up an alert on PubMed were met by an internal error, so a rerun of the search will be done towards the end of coding and on the last day of coding in order to catch any relevant newly-published papers. Coding is expected to be finished by the end of August, 2023.

Sample size rationale. We initially expected to take a random sample of the AMOA literature. However, upon completing the search we found that the number of papers located was appropriate and so was taken in its entirety. The final screening procedure has not been completed so the final sample size may be smaller than reported here.

Pilot coding. Articles were coded using a Google sheet codebook developed by the four coders during pilot coding. After initial discussion by Coder 1, Coder 3 and Coder 4 on what variables we were going to code, we applied a draft codebook to a selection of articles from a search of animal models of addiction literature looking at all substances (not just opioids). This meant a much larger pool of articles were available and, as such, low likelihood of seeing overlap between the pilot coding sample and the coding proper sample. After pilot coding these studies, the three coders came to discuss and refine variables and related search terms. Next, the Coder 2 joined in the ensuing round of pilot coding wherein five studies were coded. Coding proper was commenced when all four coders agreed on the responses taken from the five training articles and all coders were satisfied that the search terms were effective.

Data extraction. This coding procedure follows Hardwicke and colleagues (2021) and Chin and colleagues (2023). Each article is coded by two authors with disagreements being resolved through discussion between those coders and a third author if the coders do not agree. For multiple-study papers (or studies in which several steps throughout the experiment may have require, say, randomisation) the study is considered to have satisfied a variable if the characteristic is mentioned at least once. That is, if masking is mentioned once, the option of “Yes, masking mentioned in relation to this study” is selected.

With some practice, coding a single article takes about 6 minutes.

Coding variables. Please see the Codebook spreadsheet or Codebook guidelines for response options provided for each characteristics. All papers are searched using the terms provided in these documents. What follows are any additional instructions or nuances related to the coding process.

Original paper or replication was determined by scanning the abstract. Technical or biological replicates did not constitute a replication.

ARRIVE or other guidelines involved reading the “Animals” or “Subjects” section at the beginning of the method section. If guidelines other than ARRIVE were followed, the name was copy and pasted into the Google sheet. University guidelines were not included as we were primarily interested in more widely-used reporting guidelines.

The search terms for *preregistration* were designed to capture generic preregistration sites as well as those specific to preclinical research – preclinicaltrials.org, animalregistry.org.

If *supplementary materials* was coded yes, the content was checked and included in the coding procedure if relevant. For example, if the supplementary materials contained detailed methodology, search terms used to code the other variables were applied to that document as well as the original paper.

Data availability was more common than *script availability* in pilot coding. As such, there were more response options provided for data availability.

If a *sample size justification* was given, the type of justification was selected. If the justification was a power analyses, the magnitude and type (eg. Cohen's) of the effect size was coded as well as where that effect size came from., for example, from previous literature.

Although the *statistical corrections* search terms don't cover every possible type of correction, we were confident that the generic terms captured most other possibilities. For example, "the Sidak correction" is successfully captured by "correct".

Animal exclusion was coded in a way that separated outlier exclusion from exclusion due to other reasons, for example unsuccessful catheter insertion or pre-existing chamber preference in a conditioned place preference paradigm. The reasons for this are discussed in the introduction.

Lastly, we counted the *number of statistical tests* and checked for any calculation errors in the associated *p*-values. This was done by entering the entire paper's text into statcheck. The number of tests was counted by exporting the output into an Excel spreadsheet. Incorrect calculations were counted and recorded with errors indicating a decision error and those not being counted separately.

Inclusion criteria: is an empirical study; Is written in English; was published between January 2019 and August 2023; relates to opioid use including in a analgesic setting;

experiment was carried out on non-human animals; experiment contains a behavioural component (and this is mentioned in the abstract); *in vivo* research

Exclusion criteria: study investigates an opioid receptor or opioid receptor agonist but in the context of a drug of abuse other than opioids; not an empirical study, for example a literature review or meta-analysis

Data analysis. Percentages will be used to describe the prevalence of the characteristics of interest.

We also plan to conduct a test of equivalence to compare rates of open science practices with the findings of Adewumi and colleagues (2021) and Hardwicke and colleagues (2018; 2020). Additional comparisons to relevant findings may be made.

These comparisons are not confirmatory tests of a priori hypotheses. Instead, they are exploratory and, given that the comparisons are across different subdisciplines, should be interpreted with caution.

References

Adewumi, M. T., Vo, N., Tritz, D., Beaman, J., & Vassar, M. (2021). An evaluation of the practice of transparency and reproducibility in addiction medicine literature. *Addictive Behaviors*, *112*, 106560. https://doi.org/10.1016/j.addbeh.2020.106560

Chin, J., Zeiler, K., Dilevski, N., Holcombe, A., Gatfield-Jeffries, R., Bishop, R., Vazire, S., Schiavone, S.,. (2023). The transparency of quantitative empirical legal research published in highly ranked law journals (2018–2020): An observational study. *F1000 Research*, *12*(144), 23. https://doi.org/https://doi.org/10.12688/f1000research.127563.1

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society open science*, *5*(8), Article 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). *Royal Society open science*, *7*(2), 190806-190806. https://doi.org/10.1098/rsos.190806

Appendix B

Study 1: Qualitative

We decided to downgrade the ‘qualitative analysis’ of meta-analyses to be part of the literature review. This was due, firstly, to the number meta-analyses looking at animal models of opioid addiction being smaller than expected. This meant that any takeaways from this survey would be of limited relevance to animal models of opioid addiction.

Secondly, it was unexpectedly difficulty to convert all effects to one effect type (Cohen’s *d*). This was especially true because of the inconsistent sharing of data between meta-analyses.

Thirdly, while this is valuable work, it was intended to be preliminary study, and the amount of required work exceeded the author’s time and, in some cases, ability.

As such, we decided to interpret the effect sizes – where possible – using benchmarks without converting them. This meant we were still able to get a vague sense of the size of effect sizes in animal models of addiction, although it was informed by a small sample.

Methods

Inclusion/exclusion criteria changes

Systematic reviews were excluded as they did not include effect size estimates; meta-analyses looking at humans and animals were included.

This left 7 meta-analyses that provided effect sizes. To focus on the main effect sizes of interest, we extracted the effects mentioned in the abstract.

Procedure

Of the 7 meta-analyses looking at preclinical addiction research, 27 effect sizes were extracted. 16 of these were Cohen’s d; 5 were Hedges g; 2 were unstandardised meta-regression estimates; 1 was a risk ratio of dichotomous outcomes; 1 was the average mean difference between treatment groups (see Appendices C and D).

Our technique of interpreting benchmarks was not possible – to the best of our ability – for the unstandardised meta-regression estimates or the risk ratio of dichotomous outcomes. For the average mean difference, the means from the original papers were extracted from the meta-analyses and converted to Cohen's *d* (see Appendix D).

Results

Of the 24 effect sizes extracted plus the additional 1 we converted, 3 were considered small according to traditional benchmarks, 2 were considered moderate and 20 were classified as large.

Study 2: Quantitative Sample

We were able to code all 82 papers found when the research area "pharmacology, toxicology and pharmaceutics" was included in the Scopus search. However, this meant that the papers published between 12th June (date of initial database search) and the end of coding, which we had initially planned to include, had to be excluded due to time constraints.

Appendix C

The search process used to find these reviews can be found in the original preregistration (presented in Appendix A). We were interested in extracting the effect sizes from these reviews and assessing their average magnitude. This would help us determine if animal models of addiction research typically handles moderate-large effects.

Table 17

Review of Substance Abuse Related Meta-Analyses

Journal	Year	DOI	Drugs	Animal or human	Effect description	Effect type	Effect size	Interpretation according to benchmarks
Neuropsychopharmacology	2022	https://dx.doi.org/10.1038/s41386-022-01322-4	morphine & opioid	Both	opioid-sparing effect with morphine and delta-9-THC co-administration	Average mean difference	*-0.54*	*Unclear. See Appendix B*
Neuroscience and Biobehavioral Reviews	2022	https://dx.doi.org/10.1016/j.neubiorev.2022.104661	Opiate	Rodents	The results showed a large effect of pain (g = 1.37, 95% CI 1.00–1.74, p < .001) on neuronal cell death.	Hedge’s g	1.37	Large

					higher number of neonatal pain events were significantly associated with increased neuronal cell death and increased anxiety	meta-regression unstandardised	(b = −1.18, SE = 0.43, p = .006),	-
					higher number of neonatal pain events were significantly associated with increased neuronal cell death and depressant-like behavior in rodents.	meta-regression unstandardised	(b = 1.74, SE = 0.51, p = .027)	-
					Both opiates and pain had no impact on motor function	hedges g	g = 0.26	Small
Translational Psychiatry	2016	DOI: 10.1038/tp.2016.71	Ibogaine versus any	Animals	ibogaine reduced drug SA	Cohen's d	−1.54	Large
					Ibogaine did not reduce drug-induced CPP	Cohen's d	−0.22	Small
					Both the continuous and dichotomous outcome measures showed that the administration of ibogaine caused motor impairment	Cohen's d	0.82,	Large
					(Same effect as above just measured differently) Both the continuous and dichotomous outcome measures showed that the administration of ibogaine caused motor impairment	dichotomous: RR	6.2	-
					ibogaine treatment lowered drug-induced dopamine efflux in rats, as measured with dialysate levels in the nucleus accumbens and striatum after chronic cocaine or morphine use	Cohen's d	−1.14	Large

Osteoarthritis And Cartilage	2014	10.1016/j.joca.2014.06.015	Opioid	Mice, rats, guinea pigs, and rabbits	Analgesic treatment effect (SMD) was most commonly measured between drug- and vehicle treated rats with knee OA. Meta-analysis was carried out for 102 such comparisons from 26 studies. The pooled SMD was 1.36 (95% CI = 1.15-1.57).	Cohen's d	1.36	Large
					Non-steroidal anti-inflammatory drugs (NSAIDs) were associated with smaller SMDs than opioids	Cohen's d	1.16; 1.90	Large; large
					NSAID grip strength	Cohen's d	3.96	large
					NSAID mechanically evoked pain	Cohen's d	1.32	large
					NSAID weight bearing	Cohen's d	1.1	large
					NSAIDs movement evoked pain	Cohen's d	0.31	small
					Opioids mechanically evoked pain	Cohen's d	2.31	large
					Opioids weight bearing	Cohen's d	1.45	large
					Opioids movement evoked pain	Cohen's d	1.73	large
Molecular Psychiatry	2018	10.1038/mp.2017.190	Ketamine	Rodent, human and primate brain	Acute ketamine administration in rodents is associated with significantly increased dopamine levels in the cortex (Hedge's g = 1.33, P < 0.01) compared to controls	Hedge's g	1.33, P < 0.01	large
					Acute ketamine administration in rodents is associated with significantly increased dopamine levels in the striatum (Hedge's g = 0.57, P < 0.05) compared to control conditions,	Hedge's g	0.57, P < 0.05	medium
					Acute ketamine administration in rodents is associated with significantly increased dopamine levels in the nucleus accumbens	Hedge's g	1.30, P < 0.05	large

			(Hedge's g = 1.30, P < 0.05) compared to control conditions			
European Journal Of Oral Sciences	2021	10.1111/eos.12786	in animals exposed to neuropathic pain, administration of MC4R antagonist SHU9119 significantly increased paw withdrawal threshold compared to vehicle-treated animals.	Cohen's d	1.67	large
			in animals exposed to neuropathic pain, administration of MC4R antagonist HS014 significantly increased paw withdrawal threshold compared to vehicle-treated animals.	Cohen's d	2.2	large
			in animals exposed to neuropathic pain, administration of MC4R antagonists significantly and heat withdrawal latency (HS014 SMD = 3.35, 95% CI: [0.56, 6.14], I-2 = 83%) compared to vehicle-treated animals.	Cohen's d	3.35	large
Neuroscience And Biobehavioral Reviews	2013	10.1016/j.neubiorev.2012.11.018	effect of N-methyl-d-aspartate receptor (NMDAR) and B-Adrenergic receptor (B-AR) antagonists on memory reconsolidation blockade provides a potential mechanism for ameliorating the maladaptive reward memories underlying relapse in addiction	Cohen's d	0.47	medium

Appendix D

The interpretation of an effect size in one of the meta-analyses was unclear (see Appendix C). To solve this, we traced the effect of interest back to the comparisons from the papers surveyed and calculated Cohen's *d* (Statskingdom, 2022a) These results were then interpreted using Cohen's benchmarks in the same way as the effects of Appendix A.

Table 18

Transforming the Results of to Cohen's d

Study	Morphine & THC			Morphine & Vehicle					
	Mean	SD	N	Mean	SD	N	Mean difference	Cohen's d	Interpretation
1	1.12	.09	30	1.45	.08	30	-0.33	3.88	Large
2	1.13	.18	12	1.38	.18	30	-0.25	1.39	Large
3	.39	.17	7	.38	.17	28	-0.77	4.53	Large
4	.38	.08	8	.82	.07	8	-0.44	5.85	Large
5	.44	.07	30	1.5	.08	30	-1.06	14.1	Large
6	.82	.07	96	.21	.19	120	-0.61	4.26	Large
7	.39	.07	24	.74	.06	24	-0.35	17.33	Large
Total	2.25	.73	207	6.06	.83	270	-3.81	4.88	Large
Average								-0.54	

Note. Data taken from Fig. 1: Forrest plot for meta-analysis examining the opioid-sparing effect of delta-9-THC when co-administered with morphine.

Appendix E

Codebook Instructions

General instructions

- Include any supplementary materials in your search
- We are attempting to estimate the upper bound of prevalence estimates: this means that we are coding generously
- If there are multiple studies in a paper, randomisation variable is coded as "Yes, randomisation mentioned in relation to this study" if randomisation is mentioned once.
- At the end of coding a paper, all cells should be filled (use NA or – for blank/irrelevant cells)
- If there are any details you're unsure about, enter what you consider the best answer and copy and paste relevant quotations under "additional coder remarks" to discuss during coding meetings
- Pilot coding only: please time how long it takes you to code each article. This will help estimate how many papers we are able to code/indicate how easy this process is

Coding steps and instructions

1. Open spreadsheet, open Steve Haroz's statcheck (https://statcheck.steveharoz.com/)
2. Start timer (pilot coding only)
3. Open article: copy and paste DOI into library search bar
4. Check that title matches that in the Google Sheet
5. Original_replication: is this paper primarily a replication or original research?
    - We want to focus only on papers whose main goal (or one of their main goals) is to test a previously published result.
    - Limitation: We will not include papers that involve partial replications/conceptual replications/replicate methodology of another paper

- Instructions: Read abstract and search “replicat”
- Response options:
  - *Original*
  - *Replication*

6. ARRIVE: does this paper comply with the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines?
   - Here we are interested in reporting guidelines only. That means that ethics committees and university committees are likely not of interest. If unsure, select “Other” and copy & paste name into Google Sheet.
   - Instructions: Search terms "arrive" "guide" "accordance" "protocol" "reporting"
   - Response options:
     - *Yes, statement of compliance with ARRIVE or ARRIVE checklist in supplementary materials*
     - *No mention of ARRIVE or compliance with another set of reporting guidelines*
     - *Other - Mention of compliance with another set of guidelines*
7. Name_of_guidelines_if_not_arrive: what is the name of the body that created the reporting guidelines, if not ARRIVE?
   - Instructions: Copy & paste name of guidelines. Enter “NA” or “-“ if no guidelines mentioned or ARRIVE guidelines mentioned
8. Preregistration: does this study contain a statement of preregistration?
   - Is this study prereregistered and if it is, and there is a link, does the link work?
   - Instructions: search terms "regist" "osf" "aspredict" “preclinicaltrials”
   - These search terms include names of preregistration repositories
   - Response options:
     - *No, there is no statement of preregistration*
     - *Yes, there is a statement of preregistration with a link*
     - *Yes, there is a statement of preregistration with no link*

    - *There is a statement of non-registration*
    - *This paper is a registered report*

9. Supplementary_files: does the article have any supplementary information?
    - This variable is here to ensure we are checking the supplementary files for the other target characteristics
    - Instructions: search terms "suppleme" "supporting" "appendi". Check supplementary documents for characteristics of interest.
    - Response options:
        - *Yes*
        - *Yes but link broken or absent*
        - *No*
10. Data_availability: is there a statement saying that the raw data collected in this study is available?
    - Note here that we are looking for the *raw* data, not the analysed data. This means it is highly unlikely that the raw data is in the paper itself, more likely that it's in the supplementary files or accessible via a link.
    - Instructions: search terms "data" "availab" "request" "reposit"
    - Response options:
        - *No statement regarding data availability*
        - *Yes, there is a statement that the raw data is available via link*
        - *Yes, there is a statement that the raw data is available via link but link broken*
        - *Yes, there is a statement that the raw data is available via link but link absent*
        - *There is a statement that the data is UNavailable*
        - *The data is available upon request*

11. Analysis_script_availability: is the analysis script/code used to analyse the raw data available?
    - ○ This is information that the statistical package (eg. R, SPSS) would use to analyse the data.
    - ○ Instructions: search terms "code" "syntax" "script”
    - ○ Response options:
        - *No code or syntax for analysis is available*
        - *Yes, script to run statistical analyses is provided*
        - *Upon request*
12. Masking: is masking discussed in relation to this study?
    - ○ Masking and blinding are the same thing. Recall that we accept mention of any type of masking at any stage in the experiment. One mention is enough even if it is required at several points in the study.
    - ○ Instructions: search terms “blind” “mask”
    - ○ Response options:
        - *Yes, masking mentioned in relation to this study*
        - *No masking mentioned in relation to this study*
        - *Statement of no masking used*
13. Randomisation: is randomisation discussed in relation to this study?
    - ○ Recall that we accept mention of any method of randomisation. One mention is enough even if it is required at several points in the study.
    - ○ Other ways to allocate groups (eg. based on sex) are coded as “other method”
    - ○ Instructions: search terms “random” “alloc” “assign”
    - ○ Response options:
        - *Yes, randomisation mentioned in relation to this study*
        - *Other method of group allocation mentioned*
        - *No allocation method mentioned*

        - *Statement of no randomisation*

14. N_justification: have the researchers justified their sample size? If so, what is the justification?
    - Instructions: Read “animals” or “subjects” section of paper. Search terms "sample" "plan" “priori”
    - Response options:
        - *No justification given*
        - *Power analysis/sample size planning*
        - *Past research*
        - *Practical constraints*
15. If_Y_power_analysis/sample_size_planning: If a power analysis was done to determine the sample size, what was the justification for using that effect size in the power analysis?
    - Response options:
        - *No reason provided for effect size*
        - *Past research*
        - *Benchmarks (eg. Cohen’s)*
        - *SESOI: smallest effect size of interest*
        - *NA: no power analysis run*
16. If_Y_power_analysis_effect_size_type: if power analysis/sample size has been done, what statistic is used to describe the effect size?
    - Response options:
        - *Cohen's d*
        - *Hedge's g*
        - *R*
        - *R^2*
        - *SMD*

        - *Other*
        - *Not mentioned*
        - *NA: no power analysis run*

17. If_Y_power_analysis_effect_size: if power analysis/sample size has been done, what is the effect size given?
    - Instructions:
    - Response options
        - Enter number
        - *Not mentioned*
        - *NA: no power analysis run*
18. Statistical_corrections: have the researchers made statistical adjustments for multiple comparisons?
    - Instructions: Search terms "correct" "bonf" "holm" "scheffe" "Tukey" --> others? Or is there a better way of doing this search?
    - Response options:
        - *Corrected*
        - *No mention of correction method*
19. Animal_exclusion: have the researchers excluded any animals for any reason?
    - Instructions: search terms "exclu" “outl” “sacrif” “discard”
    - Response options:
        - *No statement of outlier exclusion*
        - *Yes, animals were excluded from the study*
        - *Statement of no animal exclusion*
20. Reason_for_exclusion: why was/were the animal(s) excluded?
    - Outlier exclusion: data points or subjects excluded from analysis because they fall far from the mean.

- If unsure, copy and paste the relevant aspects into the additional comments section and bring to coder meeting
- Response options:
    - *Outlier exclusion*
    - *Other*
    - *Outlier exclusion and other reason(s)*
    - *No reason given*
    - *NA: no exclusion mentioned*

21. statcheck_rows: how many test statistics are detected by Statcheck?
    - Instructions: select all article text, copy and paste into statcheck text box. If statistics detected, download as CSV to find total number of rows (remember to deduct 1 for title row). Enter number of rows
    - Limitation: this may miss *p*-values in tables or images. It may double up on in-text & in figures.
    - Do not include supplementary files in this step.
22. #statcheck_errors_not_decision: are there any *p*-values reported as incorrect?
    - Instructions: If not-bolded "INCORRECT" shows, manually count how many times. Enter number.
    - Other response options:
        - *0* = no errors
        - *NA* = no statistics detected
23. #statcheck_decision_errors: are there any errors that would change the statistically (in)significant decision?
    - If any "**INCORRECT**" in red bold, manually count and enter number
    - Other response options:
        - *0* = no errors
        - *NA* = no statistics detected
24. Additional comments: if any areas of concern/uncertainty, please note here

25. Time: how long did you spend coding this article? (for pilot coding purposes only)

○ Instructions: Stop stopwatch and enter time

Appendix F

Reconciling Coding Discrepancies

Discrepancies between coders were reconciled in the following ways:

Discrepancies in number of *p*-values detected by statcheck: discrepancies of ± 1 were not considered discrepancies. In these cases, the larger estimate was taken. This was in keeping with our aim to code generously, as the larger number would mean a smaller fraction of erroneous *p*-values. Discrepancies that were larger than ±1 were verified in the original article. It was realised that articles accessed through different databases could yield different results in statcheck, likely due to formatting differences. In these cases, Coder 1 tried to account for each coder's response to ensure the absence of error. Where both responses could be accounted for, the larger estimate was taken, again in keeping with our generous coding method.

Discrepancies in transparency measures: As with the *p*-value estimates, there were differences in reporting of transparency measures depending on the database. Where both responses could be verified, the answer demonstrating the most transparency was selected. For example, in instances where a 'statement of no data availability' and 'no statement of data availability' were verified for an article, the former was selected as the final response. This was in keeping with our intent to capture the upper bound of the rates of transparency measures.

Discrepancies in in-text measures: For characteristics that were likely to be reported in the main body of the article (randomisation, masking, multiple comparisons, sample size calculation), the possibility of a discrepancy due to differing databases was not relevant. These differences in coding were more likely to be error, in which case Coder 1 was able to verify the presence or absence of a measure by revisiting the article, or because of ambiguity. In the latter case, verification with Coder 4 was sought.

Discrepancies in 'Name of other guidelines': Seeing as the coders do not work in the preclinical context, it was difficult to judge whether guidelines mentioned included reporting stipulations. As such, coders attempted to include papers that may include reporting guidelines by including any guidelines that did not appear to be solely about ethics. Discrepancies were solved by verifying that guidelines did not include guidance about reporting experiments and by standardising guideline names. In the end, this resulted in no other guidelines about reporting were found.

Appendix G

Table 19

All Journals of Articles in the Current Study's Sample

Journal	Number articles published	Cumulative percentage
Neuropharmacology	17	6.9
Addiction Biology	15	6.1
Neuropsychopharmacology	11	4.5
Psychopharmacology	10	4.0
Behavioural Brain Research	8	3.2
Frontiers In Pharmacology	8	3.2
Drug And Alcohol Dependence	7	2.8
Neuroscience Letters	7	2.8
Int J Mol Sci	6	2.4
Frontiers In Molecular Neuroscience	5	2.0
Journal Of Pharmacology And Experimental Therapeutics	5	2.0
Pharmacology Biochemistry And Behavior	5	2.0
Addict Biol	4	1.6
Frontiers In Behavioral Neuroscience	4	1.6
J Neurosci	4	1.6
Journal Of Psychopharmacology	4	1.6
Pharmacology, Biochemistry And Behavior	4	1.6
Acta Pharmacologica Sinica	3	1.2
Behavioural Pharmacology	3	1.2
Frontiers In Neuroscience	3	1.2
International Journal Of Neuropsychopharmacology	3	1.2
Molecular Psychiatry	3	1.2
Pain	3	1.2
Progress In Neuro-Psychopharmacology & Biological Psychiatry	3	1.2
Translational Psychiatry	3	1.2
Acs Chemical Neuroscience	2	.8
American Journal Of Drug And Alcohol Abuse	2	.8
Behav Brain Res	2	.8
Behav Pharmacol	2	.8
Behavioral Neuroscience	2	.8
Biochem Biophys Res Commun	2	.8
Brain Research Bulletin	2	.8
Eneuro	2	.8

European Journal Of Pharmacology	2	.8
Experimental And Clinical Psychopharmacology	2	.8
Frontiers In Cellular Neuroscience	2	.8
Journal Of Neuroscience Research	2	.8
Neuroreport	2	.8
Proc Natl Acad Sci U S A	2	.8
Thai Journal Of Pharmaceutical Sciences	2	.8
Acta Pharmacol Sin	1	.4
Asian Journal Of Psychiatry	1	.4
Behav Neurosci	1	.4
Biological Psychiatry	1	.4
Biomedicine And Pharmacotherapy	1	.4
Br J Pharmacol	1	.4
Brain Research	1	.4
Brain, Behavior, And Immunity	1	.4
Cell Mol Neurobiol	1	.4
Cellular And Molecular Neurobiology	1	.4
Clin Exp Pharmacol Physiol	1	.4
Drug Research	1	.4
Elife	1	.4
European Journal Of Neuroscience	1	.4
European Neuropsychopharmacology	1	.4
Experimental Neurology	1	.4
Frontiers In Synaptic Neuroscience	1	.4
Genes, Brain & Behavior	1	.4
Heliyon	1	.4
Hippocampus	1	.4
Human Vaccines And Immunotherapeutics	1	.4
Ibro Neuroscience Reports	1	.4
Int J Med Sci	1	.4
Int J Neuropsychopharmacol	1	.4
J Biol Chem	1	.4
J Clin Invest	1	.4
J Psychopharmacol	1	.4
J Trace Elem Med Biol	1	.4
Journal Of Integrative Neuroscience	1	.4
Journal Of Neurochemistry	1	.4
Journal Of Neuroscience Methods	1	.4
Journal Of Pain	1	.4
Journal Of Psychiatry And Neuroscience	1	.4
Journal Of The Experimental Analysis Of Behavior	1	.4

Journal Of Venomous Animals And Toxins Including Tropical Diseases	1	.4
Learning & Memory	1	.4
Metabolic Brain Disease	1	.4
Mol Med Rep	1	.4
Mol Psychiatry	1	.4
Molecular Medicine Reports	1	.4
Molecular Pain	1	.4
Molecules	1	.4
Nature	1	.4
Nature Protocols	1	.4
Naunyn-Schmiedeberg'S Archives Of Pharmacology	1	.4
Neurobiology Of Pain	1	.4
Neurobiology Of Stress	1	.4
Neurochemical Research	1	.4
Neurochemistry International	1	.4
Neurosci Lett	1	.4
Neuroscience	1	.4
Neurotoxicology And Teratology	1	.4
Nicotine & Tobacco Research	1	.4
Nutrients	1	.4
Peptides	1	.4
Pflugers Archiv-European Journal Of Physiology	1	.4
Pharmaceutical Research	1	.4
Pharmaceuticals	1	.4
Pharmaceutics	1	.4
Pharmacol Biochem Behav	1	.4
Pharmacol Rep	1	.4
Physiological Research	1	.4
Physiology And Behavior	1	.4
Phytomedicine	1	.4
Prog Neuropsychopharmacol Biol Psychiatry	1	.4
Progress In Neuro-Psychopharmacology And Biological Psychiatry	1	.4
Psychoneuroendocrinology	1	.4
Psychopharmacology (Berl)	1	.4
Scientific Reports	1	.4
Total	247	100.0