Epidemiology and Psychiatric Sciences

cambridge.org/eps

Editorial

Cite this article: Solmi M, Correll CU, Carvalho AF, Ioannidis JPA (2018). The role of meta-analyses and umbrella reviews in assessing the harms of psychotropic medications: beyond qualitative synthesis. *Epidemiology and Psychiatric Sciences* **27**, 537–542. https://doi.org/10.1017/S204579601800032X

Received: 31 May 2018
Accepted: 1 June 2018
First published online: 16 July 2018

Key words:
Meta-analysis; psychotropic medications; safety; umbrella review

Author for correspondence:
Marco Solmi,
E-mail: marco.solmi83@gmail.com

CAMBRIDGE UNIVERSITY PRESS

The role of meta-analyses and umbrella reviews in assessing the harms of psychotropic medications: beyond qualitative synthesis

M. Solmi[1,2,3], C. U. Correll[4,5,6], A. F. Carvalho[7,8] and J. P. A. Ioannidis[9,10,11,12]

[1]Department of Neurosciences, University of Padua, Padua, Italy; [2]University Hospital of Padua, Padua, Italy; [3]Padova Neuroscience Center, University of Padua, Padua, Italy; [4]The Zucker Hillside Hospital, Department of Psychiatry, Northwell Health, Glen Oaks, NY, USA; [5]Hofstra Northwell School of Medicine, Department of Psychiatry and Molecular Medicine, Hempstead, NY, USA; [6]Charité Universitätsmedizin, Department of Child and Adolescent Psychiatry, Berlin, Germany; [7]Centre for Addiction & Mental Health (CAMH), Toronto, Ontario, Canada; [8]Department of Psychiatry, University of Toronto, Toronto, ON, Canada; [9]Department of Medicine, Stanford Prevention Research Center, Stanford, CA, USA; [10]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA; [11]Meta-Research Innovation Center at Stanford, Stanford University, Stanford, CA, USA and [12]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA

Abstract

ὠφελέειν, ἢ μὴ βλάπτειν *(Primum non nocere)* – Hippocrates' principle should still guide daily medical prescribing. Therefore, assessing evidence of psychopharmacologic agents' safety and harms is essential. Randomised controlled trials (RCTs) and observational studies may provide complementary information about harms of psychopharmacologic medications from both experimental and real-world settings. It is considered that RCTs provide a better control of confounding variables, while observational studies provide evidence from larger samples, longer follow-ups, in more representative samples, which may be more reflective of real-life clinical scenarios. However, this may not always hold true. Moreover, in observational studies, safety data are poorly or inconsistently reported, precluding reliable quantitative synthesis in meta-analyses. Beyond individual studies, meta-analyses, which represent the highest level of 'evidence', can be misleading, redundant and of low methodological quality. Overlapping meta-analyses sometimes even reach different conclusions on the same topic. Meta-analyses should be assessed systematically. Descriptive reviews of reviews can be poorly informative. Conversely, 'umbrella reviews' can use a quantitative approach to grade evidence. In this editorial, we present the main factors involved in the assessment of psychopharmacologic agents' harms from individual studies, meta-analyses and umbrella reviews. Study design features, sample size, number of the events of interest, summary effect sizes, *p*-values, heterogeneity, 95% prediction intervals, confounding factor adjustment and tests of bias (e.g., small-study effects and excess significance) can be combined with other assessment tools, such as AMSTAR and GRADE to create a framework for assessing the credibility of evidence.

Introduction

Psychopharmacologic agents may differ in both efficacy and safety (Leucht *et al.*, 2013; Solmi *et al.*, 2017; Cipriani *et al.*, 2018). While by law all medications that make their way to the market are believed to be more efficacious than placebo (Cipriani *et al.*, 2018), they may also carry variable risks of harms (Leucht *et al.*, 2013; Cipriani *et al.*, 2018) and sometimes these risks may limit tolerability of medications. Pharmacological trials' sample sizes are estimated based on the desired power (often 0.8) to detect clinically relevant effect sizes in terms of efficacy-related outcomes, but not on outcomes related to harms, safety or tolerability. Hence, individual trials are often underpowered to inform about the overall safety and tolerability of the various psychopharmacological agents. Additionally, rarely specific rating scales are used to detect and quantify adverse effects and inferential statistics are generally reserved for efficacy outcomes. Conversely, observational studies assessing psychopharmacologic agents may include larger sample sizes, longer follow-ups and more representative samples, which may be more reflective of real-life clinical scenarios. However, this may not always hold true. Moreover, in observational studies, safety data may be poorly or inconsistently reported and methodological flaws, limitations and proneness to bias may inherently decrease their credibility. Observational data include a wide spectrum of designs with highly variable levels of rigour. For example, a few observational studies that aim to assess harms may be pre-registered at the time of a new drug approval and the data may be collected prospectively according to very meticulous definitions and data collection plans and then analysed according to the prespecified protocol. Conversely, most observational studies are entirely open to

manipulation and may suffer from poor data quality and selectively reported analyses. Combining data from several observational studies and trials with meta-analytic approaches may provide a more accurate 'big picture' of extant data on the harms of prescribing psychotropics. However, if primary sources of evidence are methodologically poor, then simply synthesising evidence may lead to misleading conclusions (Ioannidis, 2017). Furthermore, while meta-analyses are typically regarded at the highest rank of evidence, they are exponentially increasing in number, often introducing more confusion than information to the literature, due to the low methodological standards of the published meta-analyses and, even more so, their included studies (Correll *et al.*, 2017), as well as redundancy, which may limit the clinical impact and the overall contribution to scientific knowledge or progress (Ioannidis, 2016; Ioannidis, 2017). It is important to comprehensively assess evidence from meta-analyses to minimise research waste (Ioannidis, 2009*b*, 2016).

While the availability of systematic reviews and meta-analyses of harms has been rather low (Papanikolaou and Ioannidis, 2004), more recently the field has witnessed a renewed attention to the reporting of harms in single studies (Ioannidis *et al.*, 2004) and this has followed proposals to improve the standardisation of reporting in meta-analyses of harms (Zorzela *et al.*, 2016). Thus, systematic reviews and meta-analyses of harms, including meta-analyses of individual-level data, may become more prevalent in the literature in upcoming years.

This editorial provides a critical overview of several aspects to account for, when assessing the quality of evidence or when grading its credibility or certainty when focusing on harms associated with the use of psychopharmacologic agents (Table 1).

Research design

Observational and intervention studies may or may not agree on their estimates of risks of harms. Differences between estimates may in some circumstances be major, especially when absolute (i.e., non-adjusted) risks are considered (Papanikolaou *et al.*, 2006). Evidence from observational and intervention studies should thus be evaluated with different frameworks.

Somewhat stricter criteria must be applied to evidence from observational studies, as they are prone to more sources of bias as well as to several sources of confounding. For example, retrospective studies are particularly prone to recall bias, while gender, smoking, age, or ongoing treatment with various antipsychotics are typical confounding factors that may influence results from observational studies.

Moreover, the adequacy, accuracy and consistency of definitions of exposure, cases and controls need to be taken into careful consideration. For example, a positive screen for depression based on a screening tool may provide a less robust outcome than a diagnosis of a major depressive episode made according to DSM-5 criteria (i.e., through a validated structured diagnostic interview) (APA, 2013). In addition, the mere presence of depressive symptoms assessed with rating scales may not substantiate the actual presence of a major depressive episode. In analogy, a self-reported accelerated heart-beat would be less reliable than a diagnosis of tachyarrhythmia made by a physician. Similarly, a definition of controls based only on the lack of a current major depressive episode would provide a less homogeneous group than a comparison group comprising individuals with a current or lifetime history of major mental disorders.

Observational studies often attempt to establish causal inferences, but this is a notoriously challenging task. Prospective cohort studies may avoid reverse causality. For example, baseline exposure (i.e. smoking) cannot be caused by a subsequent outcome (i.e. cancer). However, methodological limitations of observational studies may preclude the establishment of firm causal inferences, whilst retrospective studies cannot even sort out the possibility of reverse causality. Mendelian randomisation studies may offer a design option that may have better chances of addressing causality. Yet Mendelian randomisation studies are not particularly well-suited to study medication harms.

On the other hand, RCTs are less prone to bias, but usually, they cannot enroll desired sample sizes compared with observational studies, at least partly due to more time-consuming assessments, stricter eligibility criteria, time and economic resources needed. Exposure, namely treatment, is by definition more straightforward in RCTs compared with observational studies. Also control groups, which may vary across RCTs and which can be active (i.e., in head-to-head trials) or placebo (Weihrauch and Gauler, 1999), are clearly defined in RCTs. However, when adverse events are an outcome of interest, the sample size of individual RCTs is often too low and these studies or even meta-analyses aiming at synthesising evidence from these studies may be underpowered. Reporting also is often highly elliptical, partial, or biased (Ioannidis, 2009*a*).

In some scenarios, evidence exists from both observational studies and RCTs and this evidence ideally could be assessed together and juxtaposed. Consistency and convergence of the evidence from studies with both designs may reassure on the validity of certain associations (Papanikolaou *et al.*, 2006). For example, some previous umbrella reviews have assessed evidence from both types of study designs (Theodoratou *et al.*, 2014; Li *et al.*, 2017).

Statistics

The use of null hypothesis significance testing using standard significance thresholds (i.e., an alpha level of 0.05) has been repeatedly criticised (Wasserstein and Lazar, 2016; Szucs and Ioannidis, 2017). As a temporising measure, recently a proposal has been made to lower significance p-value threshold to 0.005 (Ioannidis, 2018). Such a threshold may ultimately aid in the identification of more robust (i.e., 'true') findings and the dropping of less consistent and less reproducible results, which may possibly contribute to the design of more methodologically sound studies in the future. Meta-analyses and umbrella reviews may also apply such thresholds to previously published evidence from RCTs. Pooling data from several RCTs may overcome the lack of power to reach this more stringent significance levels of $p < 0.005$ for harmful outcomes. For evidence derived from observational studies, even 0.005 is likely to be a lenient threshold. There is a lack of consensus on what might be an optimal threshold (or even whether a threshold should be used), but several previous umbrella reviews have used an even stricter level of $p < 10^{-6}$. Another approach is to consider falsification endpoints to adjust the p-value threshold to the peculiarities of different fields (Prasad and Jena, 2013). In this approach, p-value thresholds are tailored to the specific research setting and even to a specific database.

Several other parameters should be accounted for when assessing the evidence from meta-analyses of observational studies and of RCTs. First, publication bias and selective reporting biases may be particularly influential. There is no statistical test with high sensitivity and specificity to assess these biases and the literature

https://doi.org/10.1017/S204579601800032X Published online by Cambridge University Press

Table 1. Factors that may be considered in the assessment of the evidence on harms outcomes of pharmacological interventions from meta-analyses of observational or interventional studies

	Meta-analyses of observational studies – categorisation	Meta-analyses of intervention studies – categorisation
Design		
Study design (Veronese *et al.*, 2018)	Mendelian randomisation studies, prospective cohort, retrospective cohort or nested case-control, case-control, cross-sectional.	Triple – Double-blinded RCTs, open-label controlled trials, single-arm (naturalistic or other) interventions
Number of events (Li *et al.*, 2017)	>1000, <1000	See discussion
Primary outcome	Are harms primary or secondary outcomes	Are harms primary or secondary outcomes
Exposure, case, and control definitions	Objective measure, Structured diagnostic criteria, scales, self-report	Objective measure, structured diagnostic criteria, scales, self-report
Statistics		
Small study effects (Egger *et al.*, 1997)	Absent, present	Absent, present
Excess of significance (Ioannidis & Trikalinos, 2007)	Absent, present	Absent, present
Effect is largest (most precise) study (Li *et al.*, 2017)	As compared with the summary effect of other studies/meta-analysis	As compared with the summary effect of other studies/meta-analysis
Heterogeneity (I2) (Higgins & Thompson, 2002)	Large if >50% (consider also 95% CI)	Large if >50% (consider also 95% CI)
95% prediction intervals (Li *et al.*, 2017)	Not including null value, including a null value	Not including null value, including a null value
p value (Bellou *et al.*, 2016)	$<10^{-6}$, $<10^{-3}$, <0.05, not significant	<0.005, <0.05, not significant
Effect size (Cohen, 1988, Correll *et al.*, 2017, Sawilowsky, 2009)	Continuous – Huge (>2.0), very large (>1.2), large (>0.8), medium (>0.5), small (>0.2), very small (>0.01) Binary – Very large >5, or <0.2, large >2 or < 0.5.	Continuous – Huge (>2.0), very large (>1.2), large (>0.8), medium (>0.5), small (>0.2), very small (>0.01) Binary – Very large >5, or <0.2, large >2 or <0.5.
Adjusted analyses (Veronese *et al.*, 2018)	Adjusted, non-adjusted	Adjusted, non-adjusted
'Quality' of single studies and meta-analyses		
Methodological quality of single studies (Schünemann *et al.*, 2013, Wells *et al.*, 2013)	New-Castle Ottawa Scale (continuous score).	Cochrane Risk of Bias tool (Low, Unclear, High).
Methodological quality of meta-analyses (Correll *et al.*, 2017, Shea *et al.*, 2009, Shea *et al.*, 2017)	AMSTAR (continuous score)	AMSTAR (continuous score), AMSTAR-2, AMSTAR-plus
Reproducibility and transparency		
Computational reproducibility (statistical software codes made available)	Yes, no	Yes, no
Public datasets available	Yes, no	Yes, no
Full protocol available before publication	Yes, no	Yes, no

AMSTAR, Assessing the Methodological Quality of Systematic Reviews; optimal information size, total number of patients included in a systematic review is less than the number of patients generated by a conventional sample size calculation for a single adequately powered trial; RCT, randomised controlled trial; small study effect, when both largest study of the meta-analysis is more conservative and publication bias is present.

is replete of misleading claims where such tests are misused and misinterpreted (Lau *et al.*, 2006; Sterne *et al.*, 2011). It is probably reasonable to use a combination of tests, such as a small-study effects test (Egger *et al.*, 1997) that evaluates whether small studies could bias (i.e., inflate) the summary effect size of a meta-analytic estimate and an excess of significance test that may evaluate whether there is an excess of observed significant (i.e., 'positive') findings in relation to expected ones (Ioannidis and Trikalinos, 2007). One may also assess whether the largest study could provide a more conservative estimate than the summary effect size (Belbasis *et al.*, 2016). Furthermore, statistical measures of heterogeneity can be assessed, e.g., with I^2 >50% indicating large heterogeneity. However, often I^2 estimates are not precise (i.e., confidence intervals are large) (Ioannidis *et al.*, 2007) and statistical heterogeneity is only modestly correlated with biological and/or clinical heterogeneity. For adverse events that are uncommon, the power to detect heterogeneity between studies may be very low. Prediction intervals should also be routinely presented in meta-analyses (IntHout *et al.*, 2016), as they also accommodate the impact of between-study heterogeneity. Finally, the magnitude of the effect size should be taken into account when moving from methodological to clinical considerations of relevance and impact.

Quality of single studies and meta-analyses

There is a factory of tools that aim to assess 'quality' of studies. None of them is perfect and quality assessments based on reported features may not reflect what actually happened during the conduct of a study (Ioannidis and Lau, 1998). Considering these caveats, quality of observational studies (both case-control and cohort

https://doi.org/10.1017/S204579601800032X Published online by Cambridge University Press

studies) can be evaluated with tools, such as the New-Castle Ottawa Scale (Wells *et al.*, 2013). For RCTs, even the term 'quality' has fallen (justifiably) into disfavour and 'risk of bias assessment' is considered more appropriate, e.g., as can be conducted with the Cochrane Risk of Bias tool (Schünemann *et al.*, 2013).

For systematic reviews, Assessing the Methodological Quality of Systematic Reviews (AMSTAR) (Shea *et al.*, 2009; Pollock *et al.*, 2017) is the most popular tool to assesses the methodological 'quality' of a systematic review and meta-analysis (both observational and interventional). However, quality is almost as intangible (or more) for meta-analyses, as it is for single trials. It has been pointed out that AMSTAR scoring relies more on the 'reporting' quality, rather than 'methodological' quality (Pollock *et al.*, 2017). Also, AMSTAR completely neglects the single studies' design, pooled effect size, or sample size. Several attempts have been made to enhance AMSTAR from a mere 'methodological' scoring to a 'clinically meaningful' assessment. For example, AMSTAR-2 (Shea *et al.*, 2017) has introduced more items, accounting for the presence of randomisation or not in the interventional studies. Second, AMSTAR-plus (Correll *et al.*, 2017) in addition to study design, accounts for sample size, effect size and presence (not only assessment) of publication bias. Again, additional caveats exist, e.g., our poor ability to judge the presence of publication bias based on reported data. Moreover, the newer versions of AMSTAR do not apply to meta-analyses of observational studies.

Reproducibility and transparency

Lack of reproducibility and transparency is a major issue that should be addressed by researchers themselves and journal editors *in primis*. Unfortunately, until now, in the vast majority of studies, the raw data or even the protocols for them were not available in public (Iqbal *et al.*, 2016). However, this is hopefully going to change in the future (Munafò *et al.*, 2017). Therefore, assessing whether raw data and protocols are available for independent re-analyses may be another dimension to consider in assessing the validity and credibility of evidence. Published re-analyses in the past have shown many major differences *v.* the original publications (Ebrahim *et al.*, 2014), but this may become less of a common problem once transparency and sharing become the norm (Naudet *et al.*, 2018). When computational components are involved, sharing of computer codes helps transparency (Stodden *et al.*, 2016).

Moving from quality assessment, grading of credibility, to making recommendations

A certain degree of overlap can be found between credibility and certainty assessment across different existing frameworks. For example, according to AMSTAR-2 (Shea *et al.*, 2017) or AMSTAR-plus (Correll *et al.*, 2017), randomisation is a higher quality criterion. On the other hand, GRADE (Schünemann *et al.*, 2013) handbook retains randomisation or blinding design of RCTs as criteria that contribute to higher certainty as opposed to observational designs. Other differences can be found across grading systems. GRADE (Schünemann *et al.*, 2013) accounts also for effect size magnitude in estimating certainty of evidence, while other frameworks do not include this component in the panel of criteria to grade credibility of evidence (Bellou *et al.*, 2016). Also, while credibility grading frameworks from Ioannidis (Belbasis *et al.*, 2016) differentiate the grading of evidence from observational studies and RCTs (Theodoratou *et al.*, 2014; Li *et al.*, 2017) applying different thresholds of aforementioned features, GRADE (Schünemann *et al.*, 2013) accounts upgrades or downgrades of evidence certainty within the same framework considering both observational and randomised trial data.

Beyond quality and credibility and (un)certainty assessment, when it comes to making recommendations, additional features need to be taken into account. First of all, the clinical relevance of any finding must be considered and this has little to do with the level of statistical significance. Small effect sizes or poor relevance of outcomes of interest may preclude any recommendation, even when it seems to be based on high quality and highly statistically significant findings. Also, recommendations should always account for benefit/risk ratio. For example, a medication that is slightly more effective than an already available medication, which has a much higher frequency of severe harms, cannot really be recommended. Also, the economic evaluation of resource allocation in relation to the socio-economic burden of the disease has to be accounted for by the main stakeholders involved.

From feasibility to perfection: what is the trade-off?

Assessment of evidence on harms of psychopharmacologic medications from observational studies or RCTs should consider the assessment of multiple aspects, including research design, statistical features, quality of single studies and meta-analyses and the reproducibility and transparency of the evidence.

The full-assessment of the comprehensive list of features mentioned in this editorial may require an in-depth assessment of the published literature and, when some information is not available in the original published articles, it could be necessary to contact authors to ask further data. The extent to which unavailable information can be retrieved can vary a lot, though. Moreover, the resources needed to 'clean' the published literature after the fact may be enormous and perfect 'cleaning' may be a utopian endeavour. In-depth efforts may need to be prioritised for effects and associations that are likely to be influential, i.e., graded highly or considered to have clinical portend. In-depth looks at the data may reveal errors that render the conclusions invalid, but often the necessary additional data for such in-depth assessments and re-analyses will not be available and might be impossible to obtain. Recording of harms can be erratic and efforts at harm attribution may add extra levels of bias. Overall, it is important to use resources wisely and try to conclude objectively whether the evidence is worth trusting and to what degree. Regardless, a systematic approach is necessary to replace narrative arbitrary reviews, or reviews of reviews without systematic approaches that can be highly subjective and thus unreliable. It is hoped that the concepts covered in this editorial and summarised in Table 1 can provide a reporting and evaluation framework that can guide research and, ultimately, enhance the quality, accuracy and robustness of research findings.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None for MS, AFC and JPAI. CUC has has been a consultant and/or advisor to or has received honoraria from: Alkermes, Allergan, Angelini, Gerson Lehrman Group, IntraCellular Therapies, Janssen/J&J, LB Pharma, Lundbeck, Medavante, Medscape, Merck, Neurocrine, Otsuka, Pfizer, ROVI, Servier, Sunovion, Takeda, and Teva. He has provided expert testimony for Bristol-Myers Squibb, Janssen and Otsuka. He served on a Data Safety Monitoring Board for Lundbeck, ROVI and Teva. He received royalties from UpToDate and grant support from Janssen and Takeda. He is also a shareholder of LB Pharma.

https://doi.org/10.1017/S204579601800032X Published online by Cambridge University Press

References

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Washington, DC: Author.

Belbasis L, Bellou V and Evangelou E (2016) Environmental risk factors and amyotrophic lateral sclerosis: an Umbrella review and critical assessment of current evidence from systematic reviews and meta-analyses of observational studies. *Neuroepidemiology* **46**, 96–105.

Bellou V, Belbasis L, Tzoulaki I, Evangelou E and Ioannidis JP (2016) Environmental risk factors and Parkinson's disease: an umbrella review of meta-analyses. *Parkinsonism Related Disorders* **23**, 1–9.

Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, Leucht S, Ruhe HG, Turner EH, Higgins JPT, Egger M, Takeshima N, Hayasaka Y, Imai H, Shinohara K, Tajika A, Ioannidis JPA and Geddes JR (2018) Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* **391**, 1357–1366.

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. (Ed. Routledge). ISBN 1-134-74270-3.

Correll CU, Rubio JM, Inczedy-Farkas G, Birnbaum ML, Kane JM and Leucht S (2017) Efficacy of 42 pharmacologic cotreatment strategies added to antipsychotic monotherapy in schizophrenia: systematic overview and quality appraisal of the meta-analytic evidence. *Journal of American Medical Association Psychiatry* **74**, 675–684.

Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ and Ioannidis JP (2014) Reanalyses of randomized clinical trial data. *Journal of America Medical Association* **312**, 1024–1032.

Egger M, Davey Smith G, Schneider M and Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.

Higgins JP and Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558.

IntHout J, Ioannidis JP, Rovers MM and Goeman JJ (2016) Plea for routinely presenting prediction intervals in meta-analysis. *British Medical Journal Open* **6**, e010247.

Ioannidis JP (2009*a*) Adverse events in randomized trials: neglected, restricted, distorted, and silenced. *Archives of Internal Medicine* **169**, 1737–1739.

Ioannidis JP (2009*b*) Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *Canadian Medical Association Journal* **181**, 488–493.

Ioannidis JP (2016) The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly* **94**, 485–514.

Ioannidis J (2017) Next-generation systematic reviews: prospective meta-analysis, individual-level data, networks and umbrella reviews. *British Journal of Sports Medicine* **51**, 1456–1458.

Ioannidis JPA (2018) The proposal to lower *p* value thresholds to .005. *Journal of American Medical Association* **319**, 1429–1430.

Ioannidis JP and Lau J (1998) Can quality of clinical trials and meta-analyses be quantified? *Lancet* **352**, 590–591.

Ioannidis JP and Trikalinos TA (2007) An exploratory test for an excess of significant findings. *Clinical Trials* **4**, 245–253..

Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D and Group C (2004) Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of Internal Medicine* **141**, 781–788.

Ioannidis JP, Patsopoulos NA and Evangelou E (2007) Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal* **335**, 914–916.

Iqbal SA, Wallach JD, Khoury MJ, Schully SD and Ioannidis JP (2016) Reproducible research practices and transparency across the biomedical literature. *PLoS Biology* **14**, e1002333.

Lau J, Ioannidis JP, Terrin N, Schmid CH and Olkin I (2006) The case of the misleading funnel plot. *British Medical Journal* **333**, 597–600.

Leucht S, Cipriani A, Spineli L, Mavridis D, Orey D, Richter F, Samara M, Barbui C, Engel RR, Geddes JR, Kissling W, Stapf MP, Lassig B, Salanti G and Davis JM (2013) Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *Lancet* **382**, 951–962.

Li X, Meng X, Timofeeva M, Tzoulaki I, Tsilidis KK, Ioannidis JP, Campbell H and Theodoratou E (2017) Serum uric acid levels and multiple health outcomes: umbrella review of evidence from observational studies, randomised controlled trials, and Mendelian randomisation studies. *British Medical Journal* **357**, j2376.

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ and Ioannidis JPA (2017) A manifesto for reproducible science. *Nature Human Behaviour* **1**, 0021.

Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D and Ioannidis JPA (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. *British Medical Journal* **360**, k400.

Papanikolaou PN and Ioannidis JP (2004) Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. *American Journal of Medicine* **117**, 582–589.

Papanikolaou PN, Christidi GD and Ioannidis JP (2006) Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *Canadian Medical Association Journal* **174**, 635–641.

Pollock M, Fernandes RM and Hartling L (2017) Evaluation of AMSTAR to assess the methodological quality of systematic reviews in overviews of reviews of healthcare interventions. *BioMed Central Medical Research Methodology* **17**, 48.

Prasad V and Jena AB (2013) Prespecified falsification end points: can they validate true observational associations? *Journal of American Medical Association* **309**, 241–242.

Sawilowsky S (2009) New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* **8**, 467–474.

Schünemann H, Brożek J, Guyatt G and Oxman A (2013) Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Available at http://gdt.guidelinedevelopment.org/app/handbook/handbook.html (Accessed 31 May 2018).

Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, Henry DA and Boers M (2009) AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiol* **62**, 1013–1020.

Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E and Henry DA (2017) AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal* **358**, j4008.

Solmi M, Murru A, Pacchiarotti I, Undurraga J, Veronese N, Fornaro M, Stubbs B, Monaco F, Vieta E, Seeman MV, Correll CU and Carvalho AF (2017) Safety, tolerability, and risks associated with first- and second-generation antipsychotics: a state-of-the-art clinical review. *Therapeutics and clinical risk management*. **13**, 757–777.

Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, Carpenter J, Rucker G, Harbord RM, Schmid CH, Tetzlaff J, Deeks JJ, Peters J, Macaskill P, Schwarzer G, Duval S, Altman DG, Moher D and Higgins JP (2011) Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal* **343**, d4002.

Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JP and Taufer M (2016) Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241.

Szucs D and Ioannidis JPA (2017) When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in Human Neuroscience* **11**, 390.

Theodoratou E, Tzoulaki I, Zgaga L and Ioannidis JP (2014) Vitamin D and multiple health outcomes: umbrella review of systematic reviews and meta-analyses of observational studies and randomised trials. *British Medical Journal* **348**, g2035.

Veronese N, Solmi M, Caruso MG, Giannelli G, Osella AR, Evangelou E, Maggi S, Fontana L, Stubbs B and Tzoulaki I (2018) Dietary fiber and health outcomes: an umbrella review of systematic reviews and meta-analyses. *American Journal of Clinical Nutrition* **107**, 436–444.

https://doi.org/10.1017/S204579601800032X Published online by Cambridge University Press

Wasserstein RL and Lazar NA (2016) The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician* **70**, 129–133.

Weihrauch TR and Gauler TC (1999) Placebo-efficacy and adverse effects in controlled clinical trials. *Arzneimittel-Forschung* **49**, 385–393.

Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M and Tugwell P (2013) The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available at http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (Accessed 31 May 2018).

Zorzela L, Loke YK, Ioannidis JP, Golder S, Santaguida P, Altman DG, Moher D, Vohra S and Group PR (2016) PRISMA harms checklist: improving harms reporting in systematic reviews. *British Medical Journal* **352**, i157.

https://doi.org/10.1017/S204579601800032X Published online by Cambridge University Press