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

M. Solmi1,2,3, C. U. Correll4,5,6, A. F. Carvalho7,8 and J. P. A. Ioannidis9,10,11,12

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

Abstract

ὁμολογεῖν, ἦ 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, 2009b, 2016).

While the availability of systematic reviews and meta-analyses of harms has been rather low (Papanikolau 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 (Papanikolau 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, 2009a).

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 (Papanikolau 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
Iheterogeneity can be assessed, e.g., with
misinterpreted (Lau more conservative and publication bias is present. 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 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.

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

<table>
<thead>
<tr>
<th>Design</th>
<th>Meta-analyses of observational studies – categorisation</th>
<th>Meta-analyses of intervention studies – categorisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design (Veronese et al., 2018)</td>
<td>Mendelian randomisation studies, prospective cohort, retrospective cohort or nested case-control, case-control, cross-sectional.</td>
<td>Triple – Double-blinded RCTs, open-label controlled trials, single-arm (naturalistic or other) interventions</td>
</tr>
<tr>
<td>Number of events (Li et al., 2017)</td>
<td>&gt;1000, &lt;1000</td>
<td>See discussion</td>
</tr>
<tr>
<td>Primary outcome</td>
<td>Are harms primary or secondary outcomes</td>
<td>Are harms primary or secondary outcomes</td>
</tr>
<tr>
<td>Exposure, case, and control definitions</td>
<td>Objective measure, Structured diagnostic criteria, scales, self-report</td>
<td>Objective measure, structured diagnostic criteria, scales, self-report</td>
</tr>
</tbody>
</table>

Statistics

| Small study effects (Egger et al., 1997) | 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 (Bellasis et al., 2016). Furthermore, statistical measures of heterogeneity can be assessed, e.g., with I² >50% indicating large heterogeneity. However, often I² 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
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 assess 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 pri- nis. 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 (Munafo 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 (Bellasis 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 endeav- our. 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 been a consultant and/or advisor to or has received honoraria from: Alkermes, Allergan, Angelini, Gerson Lehrman Group, IntraCellular Therapies, Janssen/In/8], 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.

