Vibration of effects resulting from treatment selection in mixed-treatment comparisons: a multiverse analysis on network meta-analyses of antidepressants in major depressive disorder

Constant Vinatier, Clement Palpacuer, Alexandre Scanff, Florian Naudet

OPEN ACCESS

Vibration of effects resulting from treatment selection in mixed-treatment comparisons: a multiverse analysis on network meta-analyses of antidepressants in major depressive disorder

Constant Vinatier [1], Clement Palpacuer [2], Alexandre Scanff,[1] Florian Naudet [1,3]

10.1136/bmjebm-2024-112848

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjebm-2024-112848).

[1]Univ Rennes, CHU Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, Centre d'investigation clinique de Rennes (CIC1414), F-35000, Rennes, France
[2]Groupe Hospitalier de la Region de Mulhouse et Sud Alsace, Mulhouse, France
[3]Institut Universitaire de France, Paris, France

Correspondence to:
Constant Vinatier, University of Rennes, Rennes, 35000, France; constant.vinatier1@gmail.com

To cite: Vinatier C, Palpacuer C, Scanff A, *et al*. *BMJ Evidence-Based Medicine* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjebm-2024-112848

Abstract

Objective It is frequent to find overlapping network meta-analyses (NMAs) on the same topic with differences in terms of both treatments included and effect estimates. We aimed to evaluate the impact on effect estimates of selecting different treatment combinations (ie, network geometries) for inclusion in NMAs.

Design Multiverse analysis, covering all possible NMAs on different combinations of treatments.

Setting Data from a previously published NMA exploring the comparative effectiveness of 22 treatments (21 antidepressants and a placebo) for the treatment of acute major depressive disorder.

Participants Cipriani et al explored a dataset of 116 477 patients included in 522 randomised controlled trials.

Main outcome measures For each possible treatment selection, we performed an NMA to estimate comparative effectiveness on treatment response and treatment discontinuation for the treatments included (231 between-treatment comparisons). The distribution of effect estimates of between-treatment comparisons across NMAs was computed, and the direction, magnitude and statistical significance of the 1st and 99th percentiles were compared.

Results 4 116 254 different NMAs concerned treatment response. Among possible network geometries, 172/231 (74%) pairwise comparisons exhibited opposite effects between the 1st and 99th percentiles, 57/231 (25%) comparisons exhibited statistically significant results in opposite directions, 118 of 231 (51%) comparisons derived results that were both significant and non-significant at 5% risk and 56/231 (24%) treatment pairs obtained consistent results with only significant differences (or only non-significant differences) at 5% risk. Comparisons based on indirect evidence only were associated with greater variability in effect estimates. Comparisons with small absolute values observed in the complete NMA more frequently obtained statistically significant results in opposite directions. Similar results were observed for treatment discontinuation.

Conclusion In this multiverse analysis, we observed that the selection of treatments to be included in an NMA could have considerable consequences on treatment effect estimations.

Trial registration https://osf.io/mb5dy.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ It is frequent to find contradictory network meta-analyses on the same topic, although these studies are currently considered to possess among the best evidential standards.
⇒ Analytical and methodological flexibility in pairwise meta-analyses, pooled analyses and indirect comparisons can lead to vibration of effects (measuring how far an effect estimate can change across multiple distinct analyses).

WHAT THIS STUDY ADDS

⇒ Our multiverse analysis based on a large network meta-analysis exploring antidepressant efficacy in major depressive disorder suggests that network meta-analyses are prone to considerable vibration of effects, if only via the choice of treatments to be included in the network. Whether amitriptyline is more effective than other drugs—as the conclusion of the original meta-analysis—strongly depends on the drugs and comparisons considered.
⇒ Vibration of effects can be greater for treatment comparisons based solely on indirect evidence. Statistically significant results pointing in opposite directions are more readily generated when differences between treatments are small.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Results from network meta-analyses should be critically appraised.

Introduction

Network meta-analyses (NMAs) are influential evidence synthesis tools often considered to dominate the hierarchy of evidence supporting clinical decision-making.[1] By evaluating connected networks of randomised control trials (RCTs),

NMAs draw inferences on the comparative effectiveness of many interventions that may or may not have been compared directly. NMAs provide some answers to practical questions in day-to-day clinical practice, for instance. which treatment should be prioritised when many treatments are available for the same condition.[2] This information is all the more important in fields such as psychopharmacology where 'blockbuster' drugs (eg, fluoxetine for major depressive disorders) co-exist with 'me-too' drugs, marketed despite uncertain added value. Direct evidence for comparative effectiveness is indeed all too often lacking from regulatory approvals.[3] For these reasons, NMAs have become very popular tools in Evidence-Based Medicine.

NMAs are, however, victims of their own success, as their number is rapidly expanding with extensive overlap and potential redundancy. Too often, NMAs present an incomplete and fragmented picture of the total available evidence, with certain potential reproducibility issues. It has been observed that conclusions on comparative effectiveness can vary across overlapping NMAs on the same topic,[4] suggesting that NMAs are prone to vibration of effects (VoE), which measures how far an effect estimate can vary across multiple distinct analyses.[4] The study of VoE is possible by running an extreme range of sensitivity analysis. It enables assessment of the impact of a methodological choice on results by testing all of them. Various indicators exist to assess the presence of VoE,[5] such as the Janus effect (ie, when the 1st percentile and the 99th percentiles of all possible effect estimates are in different directions) or the relative OR (ROR, which is the ratio between the 1st percentile and the 99th percentile of all possible effect estimates). Several multiverse analyses have highlighted how VoE resulting from different methodological and analytical choices can lead to divergent and antagonistic conclusions in meta-analyses, for example, for pairwise meta-analyses,[6 7] for pooled analyses of individual patient data[8] and for indirect comparisons.[7] Similar reproducibility issues are expected with NMAs since they rely on strong assumptions—for example, transitivity (similarity across the studies included) and consistency (homogeneity between direct and indirect evidence)—which are quite difficult to ascertain.[9] Because of the numerous interventions compared in NMAs, they are also prone to multiplicity issues.[9] Even basic choices such as the consideration of eligible nodes to be included in an NMA can yield different effect estimates and treatment rankings.[10] We aimed to quantify and visualise VoE arising from all possible network geometries, that is, all possible combinations of treatments included in a network meta-analysis, using a multiverse analysis approach. For this purpose, we based our investigation on a widely known NMA by Cipriani *et al* exploring the comparative effectiveness and acceptability of 21 antidepressant drugs and placebo for the treatment of adults with acute major depressive disorder.[11]

Methods

Protocol, registration and reporting

The protocol was registered on 3 August 2020, on the Open Science Framework before the start of the study (available at: https://osf.io/mb5dy). The results are presented according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis checklist[12] and its extension for network meta-analyses.[13]

Data retrieval and study selection

We re-used the dataset used in Cipriani *et al*. NMA which is openly shared on Mendeley (available at: https://data.mendeley.com/datasets/83rthbp8ys/2). Data collection has been comprehensively detailed previously.[11] Briefly, this dataset was collected up to 8 January 2016 and includes published and unpublished placebo-controlled and head-to-head double blind RCTs on 21 antidepressants (agomelatine, amitriptyline, bupropion, citalopram, clomipramine desvenlafaxine, duloxetine, escitalopram, fluoxetine, fluvoxamine, levomilnacipran, milnacipran, mirtazapine, nefazodone, paroxetine, reboxetine, sertraline, trazodone venlafaxine, vilazodone and vortioxetine) used for the acute treatment of adults with major depressive disorder. Quasi-randomised trials, incomplete trials or trials that included 20% or more participants with bipolar disorder, psychotic depression or treatment-resistant depression, or patients with a serious concomitant medical condition were not included. The dataset includes 522 RCTs involving 116 477 patients in 1199 different study arms, conducted between 1979 and 2016. All study arms evaluating the efficacy of antidepressants within the licensed dose range and the accepted/recommended dose range in the main clinical guidelines[11] were considered.

Study outcome

We explored VoE for the two different outcomes used in the NMA by Cipriani *et al*. The primary outcome was efficacy assessed using the response rate (treatment response defined by a reduction of ≥50% in the total score on a standardised observer-rated scale for depression). The secondary outcome was treatment discontinuation measured by the proportion of patients who withdrew for any reason. These outcomes were recorded as close to 8 weeks after initiation of treatment as possible and computed for all randomised patients. The response rate was imputed for 292 (24.3%) study arms, and dropouts were imputed as non-responders. In the case of multi-arm studies evaluating several doses of the same treatment for which the outcome was available, these arms were pooled.

Assessment of VoEs

NMAs were performed for each possible treatment selection derived from the 21 antidepressants and placebo (ie, we constructed all possible networks with 2, 3, etc up to 22 treatments). Among these possible networks, combinations that led to non-connected networks were excluded.

For all networks included, NMAs were performed. We collected network geometry (names of treatments, number of comparisons, number of participants treated), treatment comparisons (OR and p value) and two other metrics (Cochran's Q and I^2 index). We computed the distribution of point estimates by effect sizes (ESs) and their corresponding p values under the various analytical scenarios defined by the different network geometries. Comparisons were considered statistically significant if the ES was associated with a p value <0.05. For each comparison pair, the presence of a 'Janus effect' was investigated by calculating the 1st and 99th percentiles of the distribution of the ES.[5] A Janus effect is defined as an ES that is in the opposite direction between the 1st and 99th percentiles of the meta-analysis. It demonstrates the presence of substantial VoE. In addition, we computed the distribution of the I^2 indices, and the p values on Cochran's Q test calculated for each network meta-analysis. Heterogeneity was considered statistically significant if the p value for the Q test was <0.10.

The network meta-analyses were performed with R software (V.4.2.2 (2022-10-31))[14] netmeta package (V.2.8.2), which uses a frequentist method to perform NMAs,[15] the doParallel package (V.1.0.17)[16] and the tidyverse language (V.2.0.0).[17] A random-effect model was considered for all NMAs.

Changes to the initial protocol

In addition to the Janus effect, we described two additional parameters in order to have a more comprehensive understanding of VoE in this dataset: (1) an extreme form of the Janus effect where the two extremes exhibit statistically significant results and (2) the RORs as described by Patel *et al*,[5] calculated as the ratio of ORs at the 1st and 99th percentile, which enables quantification of variations in point estimates, even when no Janus effect is observed. An ROR value of 1 suggests the absence of VoE, whereas higher ROR values indicate a more pronounced level of vibration.[18] We explored the correlation between the ROR for treatment response and the ROR for treatment discontinuation using Spearman's rank correlation.

As an exploratory analysis using our assessment of VoE for all treatment comparisons, we decided to investigate, using either a logistic or a linear model, the associations for (1) the Janus effect, (2) the existence of statistically significant results in opposite directions and (3) the RORs with the following explanatory variables considered as possible sources of VoE in NMAs: (1) a categorial variable describing the type of available evidence for the comparison in the full network and (2) the ES of the comparison in question. The type of available evidence was defined either as the presence of direct comparison without inconsistency, the presence of direct comparison with inconsistency or indirect comparisons only. A threshold for the p value <0.10 was used to define inconsistency, from a two-sided z test comparing direct and indirect evidence determined on the most complete network.[15] In this exploratory analysis, we defined the ES of the comparison in question as the absolute value of the log OR of the most complete NMA. With this last parameter, we aimed to explore whether null results were more likely to induce a Janus effect. Because of the lack of normality of residuals in the linear model for ROR, a log transformation was applied. Following a reviewer's comment, we decided to explore whether the number of treatments included in a given network impacted the presence of VoE. This was explored by computing separately the percentage of treatment comparisons that exhibited a Janus effect (among the 231 comparisons) by subgroups of NMAs with fixed numbers of treatments (ie, NMAs of 3, NMAs of 4, NMAs of 5, ..., NMAs of 21 treatments). Additionally, we plotted VoE for comparisons between the treatments exhibiting the highest and lowest VoE (clomipramine and placebo, respectively) depending on the number of treatments in the NMA. Following another reviewer's comment, we carried out sensitivity analyses using percentiles of 10%–90% and 25%–75% to define the Janus effect.

Patient and public involvement

Patients and the public were not involved in the design, conduct, reporting, or dissemination plans for this research. This was a methodological study, and we had no established contacts with specific patient groups who might be involved in this project.

Results

Primary outcome: treatment response

Among the 4 194 281 possible NMAs, 78 027 (2%) non-connected networks were excluded, resulting in a total of 4 116 254 NMAs (see online supplemental e-Table 1). The percentage of non-connected networks decreased as the number of treatments per network increased, falling from 57% for networks of two treatments to 0% for networks with 18–22 treatments. Figure 1 and online supplemental e-Table 2 summarise the distribution of the network geometries observed for all 4 116 254 NMAs included. All treatments except milnacipran and clomipramine had direct comparisons with placebo which was the most widely represented arm (with 35 721 patients). The most frequent direct comparisons were those for paroxetine versus placebo (46 studies) and fluoxetine versus placebo (40 studies). Levomilnacipran was the only treatment represented in the network by a single comparison (vs placebo). Among the 231 pairs across the 22 treatments, 99 had direct evidence and 132 relied only on indirect evidence.

Figure 2A summarises VoE observed across the 231 treatment comparisons. After computing the 4 116 254 NMAs, we observed the presence of the Janus effect in 172/231 (74%) treatment comparisons. There were statistically significant results pointing in opposite directions for 57/231 (25%) of the comparisons, suggesting the presence of substantial VoE; 56/231 (24%) comparisons obtained consistent results with only significant differences (or only non-significant differences) at the 5% level and 118/231 (51%) comparisons obtained results with both significant and non-significant results at the 5% level across NMAs. RORs ranged from 1.01 to 5.96 with a median ROR of 1.72 (IQR: 1.03–4.83) indicating significant VoE.

Clomipramine (figure 3) was the treatment with the highest level of VoE with a Janus effect present in all comparisons except the comparison with placebo. NMAs showing statistically significant results in opposite directions were present for 10 different comparisons.

Placebo (figure 4) was the treatment with the lowest level of VoE. No Janus effect was identified for any comparisons. All NMAs identified a statistically significant superiority of antidepressants over placebo, except for clomipramine and milnacipran for which 16% and 11% of the NMAs respectively failed to identify statistically significant results.

Results for other treatments are presented in online supplemental e-Figures 1–20.

Across all NMAs assessing treatment response, the median I^2 was 31% (IQR=22%–36%) and the p value for Cochran's Q test was <0.10 for 3 353 881/4 116 254 (81%) of the NMAs. Online supplemental e-Figure 21 details vibration for these two parameters.

Secondary outcome: treatment discontinuation

For these analyses, we had to exclude six studies corresponding to 145 patients randomised in seven arms because of the absence of an event. Of the 4 194 281 possible NMAs, 72 691 (2%) non-connected networks were not included, resulting in a total of 4 121 590 NMAs (see online supplemental e-Table 1), that is, 5336 more than for treatment response. Online supplemental e-Figure 22 and e-Table 2 summarise the distribution of the network geometries observed for the 4 121 590 NMAs included. Four direct comparisons that were available for treatment response were missing for treatment discontinuation (clomipramine vs milnacipran, clomipramine vs trazodone, fluoxetine vs vilazodone and reboxetine vs venlafaxine). Conversely, there were two direct comparisons for treatment discontinuation that were absent for treatment response (amitriptyline vs bupropion and clomipramine vs placebo). Among the 231 comparisons of the 22 treatments, 97 had direct evidence and 134 relied only on indirect evidence. Figure 2B summarises VoE observed across the 231 treatment comparisons. After computing the 4 194 281 NMAs, we observed a Janus effect in 180/231 (78%) treatment comparisons. We also observed statistically significant results pointing in opposite directions for 45/231 (19%) of the comparisons; 46/231 (20%) of the comparisons were able to obtain consistent results with only significant differences (or only non-significant

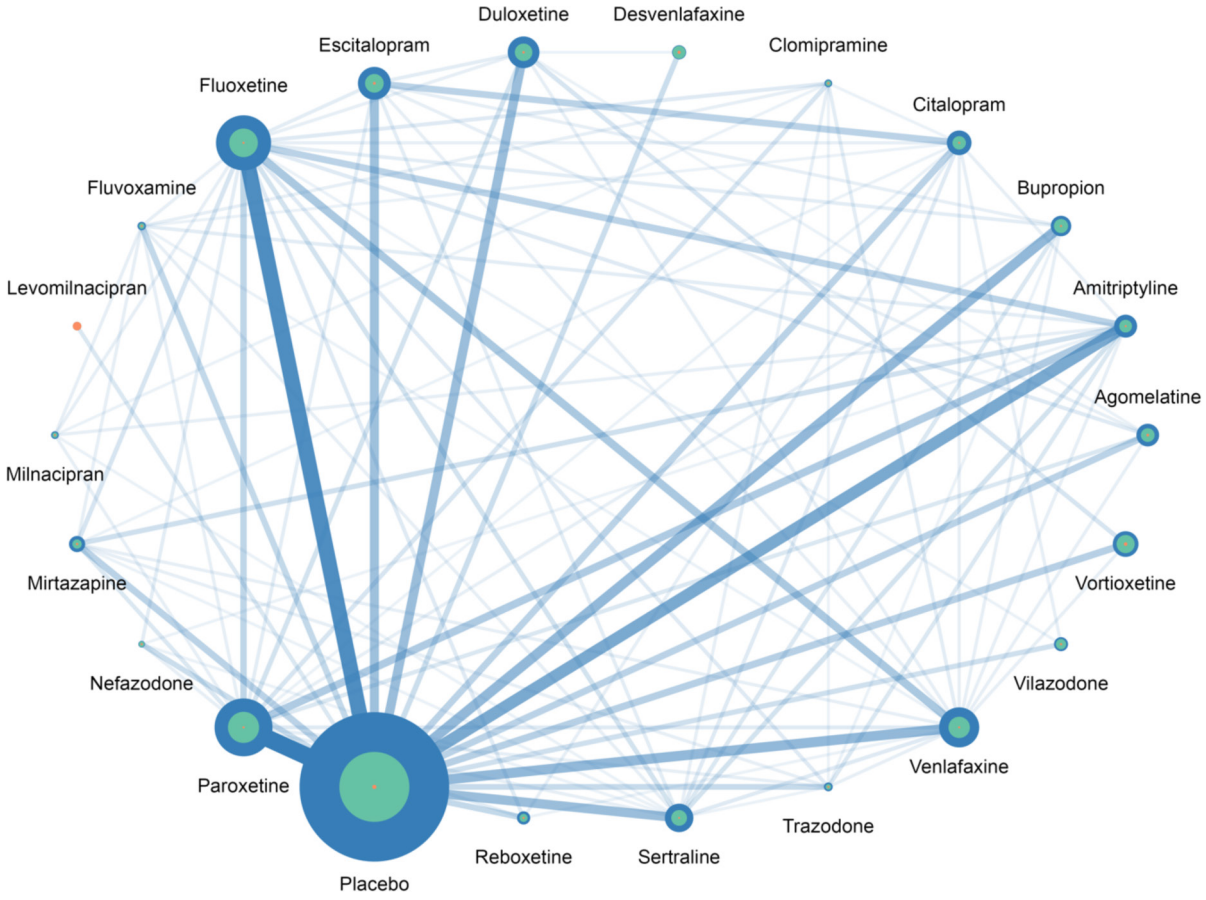

Figure 1 Distribution of network geometry for NMAs on treatment response. The size of each dot represents the number of patients allocated to the respective treatments. For each treatment, the blue circles indicate the NMAs with the largest number of patients included, the green circles represent the NMAs with the median number of patients included and the orange circles show the NMAs with the smallest number of patients included. The width of the lines is proportional to the number of trials comparing pairs of treatments in the complete NMA. NMA, network meta-analyses.

differences) at the 5% level, and 140/231 (61%) comparisons obtained results with both significant and non-significant results at the 5% level. RORs ranged from 1.01 to 10.17 with a median ROR of 1.95 (IQR: 1.33–2.50) indicating significant VoE. Results observed for all treatments are detailed in online supplemental e-Figures 23–44. Among the NMAs assessing treatment discontinuation, the median I^2 was 24% (IQR=11%–30%) and the p value on Cochran's Q test was <0.10 for 2 539 033/4 121 590 (61%) of the NMAs. Online supplemental e-Figure 45 details vibration for these two parameters. RORs observed for treatment discontinuation were correlated with RORs observed for treatment response (Spearman's ρ=0.86, p value <0.001, figure 5A).

Exploratory analysis of characteristics associated with VoE

Levomilnacipran was only studied against placebo, making it impossible to provide indirect evidence, which is why this comparison was left out for the exploratory analysis. The results based on the remaining 230 comparisons are presented in table 1. Regarding treatment response, indirect evidence was associated with a more frequent Janus effect, more results in opposite directions and greater RORs, while the ES observed in the most comprehensive meta-analysis (expressed as an absolute value of the log ORs) was only found to be associated with the Janus effect and statistically significant results in opposite directions. Quite similar results were observed for treatment discontinuation.

Exploratory analysis according to number of treatments in each network

Figure 5B details the percentage of treatment comparisons that exhibited a Janus effect (among the 231 comparisons) by subgroups of NMAs with fixed numbers of treatments. This percentage was in general above 50% for most subgroups with a maximum of 80% (180/231) for networks including nine treatments, with a gradual reduction to 30% (69/231) for meta-analyses with 21 treatments. The VoE plot for comparisons between the treatments exhibiting the highest and lowest VoE (clomipramine and placebo, respectively) according to the number of treatments in the NMA are presented in the online supplemental e-Figures 46–87.

Sensitivity analysis regarding the definition of the Janus effect

A Janus effect was identified in 58.9% (136/231) and in 35.1% (81/231) of comparisons for definitions using percentiles of 10%–90% and 25%–75%, respectively.

Discussion

Statement of principal findings

In this multiverse analysis, we performed 4 116 254 NMAs evaluating the comparative treatment response of 21 antidepressants and placebo. Depending on treatment selection, we identified substantial VoE with the presence of a Janus effect in 172/231 (74%) comparisons. For 57/231 comparisons

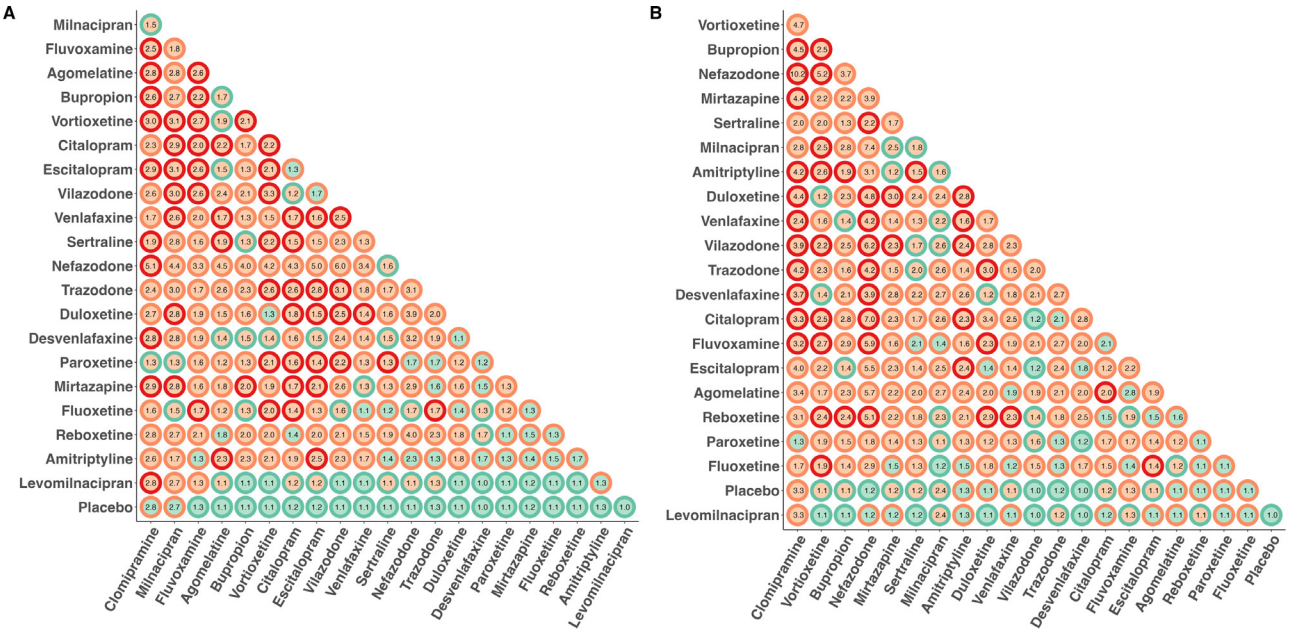

Figure 2 VoE in the 231 comparisons across the 22 treatments, classified according to their degree of VoE. (A) For treatment response. (B) For treatment discontinuation. For each dot, the centre indicates the existence of a Janus effect (green=no, orange=yes), the outline indicates the existence of statistically significant results in two opposite directions (green=comparisons that obtained consistent results with only significant differences (or only non-significant differences) at the 5% level, orange: comparison that yielded both significant and non-significant results at the 5%, red: significant results observed in opposite directions). Numbers correspond to the relative ORs which are ratios quantifying the ratios of ORs at the 1st and 99th percentile. The higher the relative OR, the greater the variability of results arising from the network geometries considered. VoE, vibration of effects.

(25%), VoE yielded statistically significant results in opposite directions. In more concrete terms, whether amitriptyline is more effective than other drugs, as suggested by Cipriani *et al*, strongly depends on the drugs and comparisons considered. Similar results were observed among the 4 121 590 NMAs evaluating treatment discontinuation. RORs for treatment response and treatment discontinuation were highly correlated. Comparisons relying on indirect evidence alone were associated with all three indices of VoE (Janus effect, significant results in opposite directions and RORs). Having an ES close to zero (as assessed in the most comprehensive meta-analysis) was associated with the Janus effect, with significant results in opposite directions, but not with RORs.

In other terms, variations in estimated effects are greater for comparisons relying on indirect evidence only. When the actual differences between treatments are small, this can

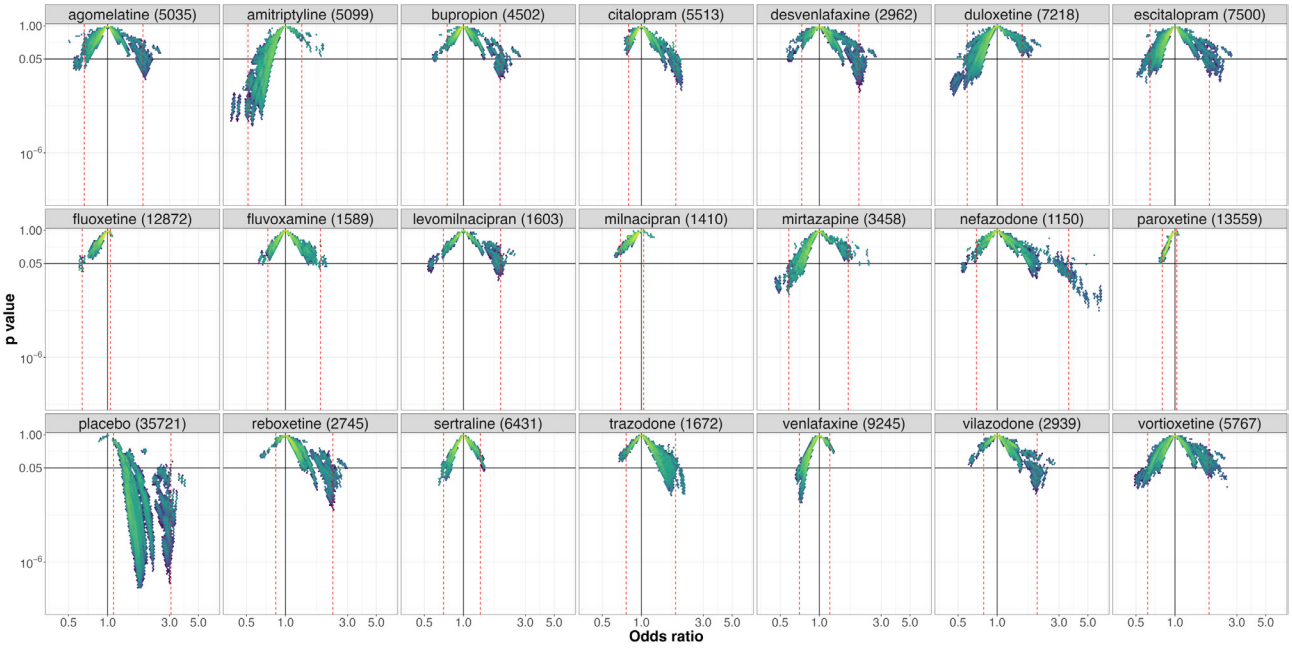

Figure 3 Vibration of effects for treatment response for the comparisons of clomipramine with the 20 remaining antidepressants and placebo (with the number of patients included in the most complete network for this comparison). An OR >1 favours clomipramine. The colours indicate the log densities of network meta-analyses (yellow: high, green: moderate, blue: low). Dotted red lines show the 1st and 99th percentiles.

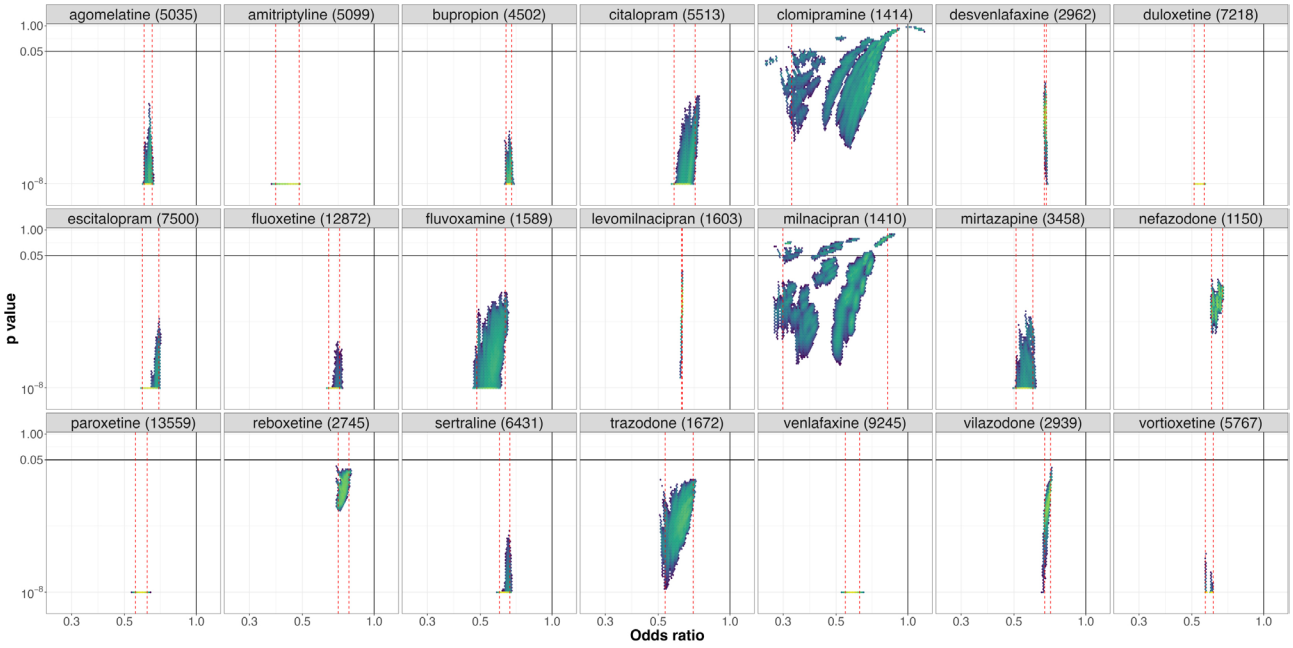

Figure 4 Vibration of effects for treatment response for the comparisons of placebo with the 21 antidepressants (with the number of patients included in the most complete network for this comparison). An OR >1 favours placebo. The colours indicate the log densities of network meta-analyses (yellow: high, green: moderate, blue: low). Dotted red lines show the 1st and 99th percentiles.

lead to effect estimates in opposite directions, and occasionally to statistically significant results in both directions. It is not surprising to see these results in this very specific multiverse analysis focused on antidepressants. Many of the drugs studied are me-too drugs from a few therapeutic classes, resulting in small difference between treatments. In addition, VoE could be expected in this corpus, as a previous re-analysis of the Cipriani *et al* dataset was able to identify differences among antidepressant placebos although all are composed of sucrose,[19] to some extent suggesting violations of the main assumption of NMAs.

Strengths and weaknesses of the study

We used a well-known NMA with 22 different treatments (including placebo), making it possible to study a large number of network geometries. As this was a multiverse analysis performed in a very specific field (the use of antidepressants to treat major depressive disorder), different results could be observed in a different

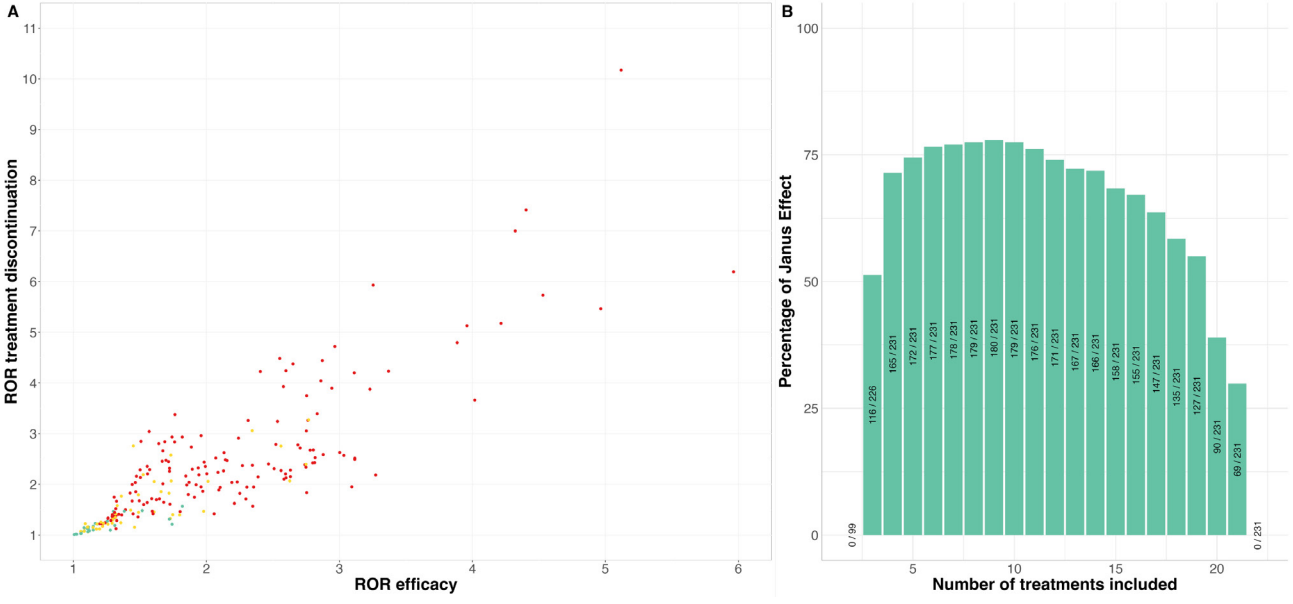

Figure 5 Post-hoc analysis. (A) RORs for treatment response and treatment discontinuation. The colour of the dots indicates the presence of a Janus effect for both outcomes (green: no Janus effect, yellow: Janus effect for one of the outcomes and red: Janus effect for both). (B) Percentage of treatment comparisons exhibiting a Janus effect (among the possible comparisons) by subgroups of NMAs with fixed number of treatments (ie, NMAs of 3, NMAs of 4, NMAs of 5, ..., NMAs of 21 treatments). NMA, network meta-analyses; ROR, relative OR.

Table 1 Association between vibration of effects indices and various characteristics of treatment comparisons

Characteristics of treatment comparisons	Janus effect		Statistically significant results in opposite directions		Log (relative OR)	
	OR (95% CI)	P value	OR (95% CI)	P value	β (95% CI)	P value
Treatment response						
Type of available evidence for the comparison (ref=direct evidence without inconsistency)						
Direct evidence with inconsistency	0.393 (0.149 to 1.016)	0.055	0.596 (0.126 to 2.117)	0.458	0.041 (−0.109 to 0.190)	0.593
Only indirect evidence	3.176 (1.587 to 6.492)	0.001*	2.225 (1.109 to 4.699)	0.029*	0.406 (0.313 to 0.498)	<0.001*
Effect size*	0.037 (0.004 to 0.373)	0.005*	0.004 (0.000 to 0.092)	0.002*	−0.072 (−0.396 to 0.252)	0.660
Treatment discontinuation						
Type of available evidence for the comparison (ref=direct evidence without inconsistency)						
Direct evidence with inconsistency	0.643 (0.221 to 1.943)	0.4217	0.906 (0.128 to 4.113)	0.907	1.134 (0.941 to 1.366)	0.184
Only indirect evidence	2.773 (1.281 to 6.141)	0.0103*	3.641 (1.600 to 9.433)	0.004*	1.632 (1.456 to 1.828)	<0.001*
Effect size*	3.25×10^{-6} (2.68×10^{-8} to 2.56×10^{-4})	<0.001*	0.063 (0.001 to 4.329)	0.214	0.667 (0.350 to 1.272)	0.217

*p-value <0.05

*Effect sizes are expressed as absolute values of the log relative OR estimated in the most complete network meta-analysis. Relative ORs quantify the ratio of ORs at the 1st and 99th percentile. The higher the relative OR, the greater the variability of results arising from the network geometries considered.

context, for example, for networks of different size or in different fields. In addition, estimating VoE related to treatment selection could be difficult to conduct for NMAs exploring smaller networks of RCTs. Smaller networks could be less prone to VoE because of the treatment selection, as the contribution of indirect comparisons is associated with the number of treatments included in the NMA.[20] On the other hand, studies of VoE related to treatment selection can be challenging in larger networks, since performing a large number of NMAs requires a lot of computing time. It took us almost 3 months on a personal computer to compute the near 8 million NMAs needed for this specific case study.

In addition, we considered only treatment selection as a source of VoE for this study. Although it seems to be a relevant choice, as differences in treatment selection are frequently observed for overlapping NMAs on the same topic,[4] complementary methodological choices could have been made, for example, the exhaustiveness of the evidence base (related both to the selection criteria and to the quality of the literature searches) or the risk of bias in the RCTs included. In addition, for treatment selection, additional VoE could be related to decisions made to merge or not to merge different doses of the same treatment in a given node. Lastly, the exploratory analysis of the characteristics associated with VoE includes results on various treatments that are in fact correlated, meaning that uncertainty is greater than that reflected by the 95% CIs. Great caution is therefore warranted when interpreting these exploratory results.

Strengths and weaknesses in relation to other studies, discussion of important differences in results

After our previous multiverse analysis, which made several methodological choices for indirect comparison meta-analyses to compare nalmefene and naltrexone in the reduction of alcohol consumption,[7] this new study, in a more complex network, corroborates VoE arising from indirect comparisons. VoE was also found to influence the results in a head to head meta-analysis in the case of acupuncture for smoking cessation, a domain that is known for its clinical and methodological heterogeneity.[21] Similarly, marked VoE was observed in a meta-analysis comparing operative with non-operative treatments for proximal humerus fractures.[6] While the domain of antidepressant research is probably more standardised with less variability in interventions and study designs than acupuncture or surgery, we were still able to find evidence for VoE. In addition, VoE has been observed in pooled analyses of individual participant data from 12 RCTs comparing canagliflozin and placebo for type 2 diabetes mellitus.[8] All these multiverse analyses were useful to investigate reproducibility issues and controversies arising from redundant and overlapping meta-analyses.[22] Nevertheless, these studies converge to point to the existence of VoE in meta-analyses, and we recommend further research to systematically explore VoE and its determinants (eg, ESs, heterogeneity, inconsistency, risk of bias in studies included and random sampling) in a large set of meta-analyses before any systematic implementation in routine practice. It might help to understand better the strengths and limitations of the approach, even if computational time can be a source of difficulties.

Meaning of the study: possible explanations and implications for clinicians and policymakers

Our results show that effect estimates in NMAs can be impacted by the network structure. In other words, NMAs allow for a certain amount of analytical flexibility, which can lead to divergent results, and NMAs can therefore be easily hijacked to a desired conclusion. This is all the more important since NMAs have particular importance for clinical decision-making: since direct evidence of comparative effectiveness is all too often lacking in regulatory approvals,[23] indirect evidence is often required for guideline development.[24] Concerning the conduct of NMAs, analytical flexibility can be partly addressed by pre-registration in Prospero,[25] a practice that is encouraged but not enforced by most journals, as there is no policy for meta-analyses similar to the 2005 ICMJE policy on clinical trials.[26] Still, because meta-analyses are almost always retrospective studies that gather existing evidence, the possibility of an a posteriori registration is often very difficult to rule out. The constitution of systematic, permanent, living NMAs could also help to reduce reporting bias of this sort. Regarding interpretation of NMAs, our results highlight the importance of considering uncertainties in NMA results, and corroborate the widespread idea that indirect comparisons can lead to biased conclusions.[27 28] This is all the more important since empirical evidence suggests that in NMAs, most of the information often comes from indirect evidence.[20] NMAs results are considered as an important source of evidence for clinical practice guidelines,[24] for instance, in mental health disorders.[29] However, our results raise doubts about the relevance of indirect comparisons as a decision-making tool, and provide empirical support for the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) approach for NMAs, which considers the certainty of evidence for all direct, indirect and NMA estimates

between interventions included in the network,[30] and downgrades certainty of the evidence in case of absence of direct comparisons. In the case of antidepressants for major depressive disorder, achieving a precise classification of antidepressants is challenging. Only 18% of the clinical trials were rated by Cipriani *et al* as having a low risk of bias.[11] Our multiverse analysis suggests that the inclusion of different treatments in the network adds even more uncertainty and that particular caution should be exercised when ranking treatments.

Unanswered questions and future research

In this multiverse analysis, we explored the VoE arising from the treatment selection in a large NMA on 21 antidepressants and placebo in the treatment of major depressive disorders. We found substantial variations in the magnitude, direction and statistical significance of the effects estimated. These findings suggest that when conducting NMAs on RCTs, the selection of treatments to be included in the network could have considerable consequences on treatment effect estimations. More comprehensive studies on VoE across the medical literature are needed to gain better understanding of these reproducibility issues and to define safeguards to limit their impact on clinical decision-making.

X Florian Naudet @NaudetFlorian

Acknowledgements We thank Angela Verdier for revising the English, and Karima Hammas for her help for the protocol. We would like to express our gratitude to the FABrique du Loch for granting us access to a calculation server, which has contributed to our efforts to reduce computation time.

Contributors CP initiated and designed the study. FN initiated and designed the study, interpreted the results and drafted the manuscript. CV cleaned the data, performed the analysis, interpreted the results and drafted the manuscript. AS contributed to the data analysis. All authors have critically revised the manuscript for important intellectual content and approved the manuscript. FN acts as guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The data and code are openly shared on the OSF (https://osf.io/hb7uj/) and the dataset is shared by the original authors on Mendeley (https://data.mendeley.com/datasets/83rthbp8ys/2).

ORCID iDs
Constant Vinatier http://orcid.org/0000-0002-6899-1838
Clement Palpacuer http://orcid.org/0000-0001-5440-1860
Florian Naudet http://orcid.org/0000-0003-3760-3801

References

1 **Leucht S**, Chaimani A, Cipriani AS, *et al*. Network meta-analyses should be the highest level of evidence in treatment guidelines. *Eur Arch Psychiatry Clin Neurosci* 2016;266:477–80.
2 **Mills EJ**, Thorlund K, Ioannidis JPA. Demystifying trial networks and network meta-analysis. *BMJ* 2013;346:f2914.
3 **Erhel F**, Scanff A, Naudet F. The evidence base for psychotropic drugs approved by the European medicines agency: a meta-assessment of all European public assessment reports. *Epidemiol Psychiatr Sci* 2020;29:e120.
4 **Naudet F**, Schuit E, Ioannidis JPA. Overlapping network meta-analyses on the same topic: survey of published studies. *Int J Epidemiol* 2017;46:1999–2008.
5 **Patel CJ**, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* 2015;68:1046–58.
6 **Sandau N**, Aagaard TV, Hróbjartsson A, *et al*. A meta-epidemiological study found that meta-analyses of the same trials may obtain drastically conflicting results. *J Clin Epidemiol* 2023;156:95–104.
7 **Palpacuer C**, Hammas K, Duprez R, *et al*. Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Med* 2019;17:174.
8 **Gouraud H**, Wallach JD, Boussageon R, *et al*. Vibration of effect in more than 16 000 pooled analyses of individual participant data from 12 randomised controlled trials comparing canagliflozin and placebo for type 2 diabetes mellitus: multiverse analysis. *BMJ Med* 2022;1:e000154.
9 **Faltinsen EG**, Storebø OJ, Jakobsen JC, *et al*. Network meta-analysis: the highest level of medical evidence? *BMJ Evid Based Med* 2018;23:56–9.
10 **Mills EJ**, Kanters S, Thorlund K, *et al*. The effects of excluding treatments from network meta-analyses: survey. *BMJ* 2013;347:f5195.
11 **Cipriani A**, Furukawa TA, Salanti G, *et al*. 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* 2018;391:1357–66.
12 **Page MJ**, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
13 **Hutton B**, Salanti G, Caldwell DM, *et al*. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med* 2015;162:777–84.
14 **R Development Core Team**. R: a language and environment for statistical computing. Vienna R Foundation for Statistical Computing; 2009.
15 **Balduzzi S**, Rücker G, Nikolakopoulou A, *et al*. netmeta: An R package for network meta-analysis using frequentist methods. *J Stat Softw* 2023;106:1–40.

16 **Daniel F**, Corporation M, Weston S, *et al*. doParallel: Foreach parallel Adaptor for the « parallel » package. 2022.
17 **Wickham H**, Averick M, Bryan J, *et al*. Welcome to the Tidyverse. *JOSS* 2019;4:1686.
18 **Klau S**, Patel CJ, *et al*. Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. *MP* 2023;7.
19 **Holper L**, Hengartner MP. Comparative efficacy of placebos in short-term antidepressant trials for major depression: a secondary meta-analysis of placebo-controlled trials. *BMC Psychiatry* 2020;20:437.
20 **Papakonstantinou T**, Nikolakopoulou A, Egger M, *et al*. In network meta-analysis, most of the information comes from indirect evidence: empirical study. *J Clin Epidemiol* 2020;124:42–9.
21 **El Bahri M**, Wang X, Biaggi T, *et al*. A multiverse analysis of meta-analyses assessing acupuncture efficacy for smoking cessation evidenced vibration of effects. *J Clin Epidemiol* 2022;152:140–50.
22 **Ioannidis JPA**. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses: mass production of systematic reviews and meta-analyses. *Milbank Q Sept* 2016;94:485–514.
23 **Vokinger KN**, Glaus CEG, Kesselheim AS, *et al*. Therapeutic value of first versus supplemental indications of drugs in US and Europe (2011-20): retrospective cohort study. *BMJ* 2023;382:e074166.
24 **Kanters S**, Ford N, Druyts E, *et al*. Use of network meta-analysis in clinical guidelines. *Bull World Health Organ* 2016;94:782–4.
25 **Page MJ**, Shamseer L, Tricco AC. Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Syst Rev* 2018;7:32.
26 **De Angelis C**, Drazen JM, Frizelle FA, *et al*. Clinical trial registration: a statement from the International committee of Medical Journal editors. *N Engl J Med* 2004;351:1250–1.
27 **Jansen JP**, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med* 2013;11:159.
28 **Nikolakopoulou A**, Higgins JPT, Papakonstantinou T, *et al*. Cinema: an approach for assessing confidence in the results of a network meta-analysis. *PLOS Med* 2020;17:e1003082.
29 **Malhi GS**, Bell E, Bassett D, *et al*. The 2020 Royal Australian and New Zealand college of psychiatrists clinical practice guidelines for mood disorders. *Aust N Z J Psychiatry* 2021;55:7–117.
30 **Izcovich A**, Chu DK, Mustafa RA, *et al*. A guide and pragmatic considerations for applying GRADE to network meta-analysis. *BMJ* 2023;381:e074495.