BMJ Mental Health

Crossroads of methodological choices in research synthesis: insights from two network meta-analyses on preventing relapse in schizophrenia

Giovanni Ostuzzi ,[1] Johannes Schneider-Thoma ,[2] Federico Tedeschi,[1] Stefan Leucht ,[2] Corrado Barbui[1]

[1]WHO Collaborating Centre for Research and Training in Mental Health and Service Evaluation, Department of Neuroscience, Biomedicine and Movement Sciences, Università degli Studi di Verona, Verona, Italy
[2]Department of Psychiatry and Psychotherapy, School of Medicine, Technical University of Munich, Munich, Germany

Correspondence to
Dr Giovanni Ostuzzi, WHO Collaborating Centre for Research and Training in Mental Health and Service Evaluation, Department of Neuroscience, Biomedicine and Movement Sciences, University of Verona, Verona, 37129, Italy; giovanni.ostuzzi@univr.it

GO and JS-T are joint first authors.
SL and CB are joint senior authors.

Received 10 February 2023
Accepted 10 May 2023
Published Online First
17 May 2023

To cite: Ostuzzi G, Schneider-Thoma J, Tedeschi F, *et al*. *BMJ Ment Health* 2023;**26**:1–4.

ABSTRACT

In recent years, network meta-analyses have been increasingly carried out to inform clinical guidelines and policy. This approach is under constant development, and a broad consensus on how to carry out several of its methodological and statistical steps is still lacking. Therefore, different working groups might often make different methodological choices based on their clinical and research experience, with possible advantages and shortcomings. In this contribution, we will critically assess two network meta-analyses on the topic of pharmacological prevention of relapse in schizophrenia, carried out by two different research groups. We will highlight the implications of different methodological choices on the analysis results and their clinical–epidemiological interpretation. Moreover, we will discuss some of the most relevant technical issues of network meta-analyses for which there is not a broad methodological agreement, including the assessment of transitivity.

Pharmacological prevention of relapse in schizophrenia is a debated topic. Recently, two different research groups attempted to address this relevant clinical issue by synthesising data from randomised controlled trials (RCTs) using network meta-analysis (NMA) methodology.[1 2] In this contribution, jointly written by some of the researchers involved in both NMAs, similarities and differences of these two approaches are examined in order to discuss how different methodological choices can be applied to the same clinical problem and whether they contributed to differences in results and conclusions. As a second aim, we critically appraised some of the current technical issues of NMAs, exemplified in these two NMAs, suggesting possible future developments.

SIMILARITIES AND DIFFERENCES BETWEEN THE NMAS

Both NMAs included RCTs enrolling clinically stable adults with schizophrenia spectrum disorders (table 1), relying mostly on the definition of 'clinical stability' provided by primary studies, but with slight differences. Ostuzzi *et al*[2] additionally included those RCTs where, although individuals were not clearly defined as 'clinically stable', mean severity scores (eg, Brief Psychiatric Rating Scale) at baseline indicated relatively low levels of psychopathology, according to commonly employed cut-offs. This led to the inclusion of 14 RCTs (1486 individuals) that would have been otherwise excluded. Further, Schneider-Thoma *et al*[1] excluded individuals with prominent negative symptoms, considering that this specific subgroup of individuals may not be comparable (transitive) with the target population in different regards, while Ostuzzi *et al* did not apply this exclusion criterion. This accounted for some differences, such as amisulpride not being included in the network of Schneider-Thoma *et al*, as it was only investigated in trials for prominent negative symptoms. Overall, for the primary outcome (ie, relapse), Ostuzzi *et al* included 89 RCTs (22 275 participants) and Schneider-Thoma *et al* included 100 RCTs (16 812 participants). Despite similar inclusion criteria, only 44 RCTs contributed to this analysis for both NMAs.

We note that being more or less inclusive ultimately depends on methodological considerations, as well as the clinical perspective of the working group. On one hand, being inclusive might increase heterogeneity between studies and threaten the assumption of transitivity, which postulates that included RCTs should be similar in the distribution of all potential effect modifiers except for the treatments being compared.[3] On the other hand, a more inclusive approach might have the benefits of increasing the statistical power and the connectivity of the network, improving external validity and applicability of results by including a broader range of clinical features commonly seen in real-world practice.

Moreover, employing rating scales' cut-off scores to include/exclude RCTs in meta-analyses is not a routine approach and could be questioned. In this case, 'clinical stability' is a complex construct that might not be exhaustively informed by symptom severity only. By contrast, as RCTs employ heterogeneous definitions of 'clinical stability', using a common cut-off could be a more consistent measure of symptom severity across different studies.

Similar considerations can apply to the construct of 'relapse' and how it should be measured. Both NMAs used the definition of relapse provided by the authors of primary studies, but Ostuzzi *et al* additionally imputed missing data from mean change scores, again using cut-off thresholds (ie, an increase of at least 25% of the baseline Positive and Negative Syndrome Scale at the end of the study). Although such thresholds have been

Table 1 Main differences between the two network meta-analyses

	Schneider-Thoma *et al*,[1] *Lancet* 2022	Ostuzzi *et al*,[2] *World Psychiatry* 2022
Population	► Adults with schizophrenia spectrum disorders ► Clinically stable (as reported by primary studies) ► Individuals with prominent negative symptoms were excluded	► Adults with schizophrenia spectrum disorders ► Clinically stable (as reported by primary studies or, alternatively, according to the mean baseline severity score)
Intervention/comparison	All available oral and LAI second-generation antipsychotics as well as a selection of 19 first-generation antipsychotics	All available oral and LAI antipsychotics
Outcome	Relapse (as reported by primary studies)	Relapse (as reported by primary studies; if not available, imputed from mean change in severity score)
Studies included	RCTs	RCTs enrolling 50 or more participants
Design	Bayesian network meta-analysis (using ORs as effect size measure for analysis; for presentation of results, ORs transformed to risk ratios to increase interpretability)	Frequentist network meta-analysis (using risk ratios as effect size measure)
Transitivity assessment	Visual inspection of box plots	Visual inspection of box plots; Kruskal-Wallis test and meta-regression analysis
Risk of bias assessment	Cochrane Risk of Bias 2	Cochrane Risk of Bias 2
Quality assessment	CINeMA	CINeMA
Conclusions	Quote: 'As we found no clear differences between antipsychotics for relapse prevention, we conclude that the choice of antipsychotic for maintenance treatment should be guided mainly by their tolerability.'	Quote: 'Based on these findings, olanzapine, aripiprazole and paliperidone are the best choices for the maintenance treatment of schizophrenia spectrum disorders, considering that both LAI and oral formulations of these antipsychotics are among the best-performing treatments and have the highest confidence of evidence for relapse prevention.'

CINeMA, Confidence in Network Meta-Analysis; LAI, long-acting injective antipsychotic; OR, Odds Ratio; RCTs, randomised controlled trials.

commonly used as an approximation of relapse in randomised trials,[4] there is debate around the definition of clinically relevant change according to rating scale scores, and some cut-offs might be appropriate for some subpopulations of patients (ie, negative symptoms) and not for others.[4–6]

Notably, possible biases related to broad inclusion criteria and data imputation can be tested by means of sensitivity analyses, that is, excluding RCTs with specific assumptions (for example, those for which 'clinical stability' or 'relapse' was imputed). In this case, sensitivity analyses performed by both working groups were largely consistent with primary results, supporting the pragmatism of such methodological choices.

Of relevance, both NMAs compared individual antipsychotics against each other; however, the same antipsychotics were used in the context of different study designs, reflecting different treatment strategies. In particular, participants stabilised with one antipsychotic might be randomised to continuing, decreasing the dose, switching or stopping it (ie, switch to placebo). This might bias the interpretation of results, as different treatment strategies might be included under the same antipsychotic (and therefore the same 'node').[7]

CURRENT TECHNICAL ISSUES EXEMPLIFIED IN THE TWO NMAS

Despite growing literature and guidelines on how to technically carry out an NMA, many choices are ultimately taken on the basis of very pragmatic considerations, including different perceptions on the nature of the clinical problem, its application in real-world practice, as well as clinical and research experience. This should be regarded as a value because, as long as the methodology is preplanned, transparently and rigorously applied, it allows different viewpoints of the same clinical phenomenon, giving nuances to the discussion around the practical application of clinical–epidemiological data.

Although NMAs are increasingly carried out, and international institutions (such as the Cochrane Collaboration)[8] and experts are constantly updating methodological guidelines, some technical and practical issues have not been standardised yet. We chose to discuss some of those that are practically exemplified in these two NMAs.

First, although transitivity is an essential assumption of NMAs, standardised approaches to systematically assess (and ideally quantify) clues of its violation are not available. Many published NMAs do not even perform a formal assessment of this assumption. According to the Cochrane Handbook: 'transitivity can be evaluated by comparing the distribution of effect modifiers across the different comparisons',[8] which is usually done by visually inspecting box plots of effect modifiers by treatment edges. Moreover, given that the current method of choice for testing for global inconsistency is the design-by-treatment interaction model,[9] built on the idea that inconsistency may not only be at the 'loop', but also at the 'design' level (ie, the list of compared treatments), box plots by design of the study is another valid approach.[10] Comparing treatment edges might be more convenient if causes of possible inconsistency are sought, as the core of NMAs is the identification of treatments effects, which are calculated through pairwise comparisons of outcomes. On the other hand, using study designs allows for testing the homogeneity of distribution across possible effect modifiers, which might be biased when using treatment edges in the presence of multiarm studies. More in general, we note that an accurate interpretation of box plots is rather difficult and subjective, particularly when the network is sparse (ie, with many comparisons and only few trials for comparison). Aiming to overcome these difficulties, Ostuzzi *et al* attempted to adopt an inferential approach to detect possible distribution imbalances (ie, a Kruskal-Wallis test), on the grounds that, in a random-effects model, the assumption of equal distribution of effect modifiers across comparisons does not hold exactly (that would not be realistic, due to absence of randomisation between trials), but only in expectation.[11] Even such approach however does not give a definitive answer on the possible causes of inconsistency, since tests across nodes or edges would not respect the independence assumption, while tests across designs are likely to suffer from lack of power unless there is a high number of trials for each design (which is very unlikely). However, it should be highlighted that absence of inconsistency should not be interpreted as evidence of transitivity, whose

assumption should be assessed on a theoretical level before the NMA is conducted.[12]

Second, Ostuzzi *et al* excluded small studies (including less than 50 participants), considering that they tend to show higher heterogeneity and they may bias estimates due to omission of publications with non-significant findings in case of publication bias.[13] However, as noticed by Zhang *et al*, omission of studies from NMAs, regardless of the reason, may have a substantial impact on estimated results[14] and is therefore not recommended in general.[15] Considering that sample size did not act as an effect modifier, inclusion of small studies could have led to higher precision of estimates in Ostuzzi *et al*. However, it is also possible that it could have led to higher heterogeneity and subsequently reduced precision in the random-effects NMA model as it is the case in Schneider-Thoma *et al*.

Third, the two working groups used different approaches to the analysis, namely Bayesian (Schneider-Thoma *et al*) and frequentist approach (Ostuzzi *et al*). As noted by Seide *et al*, the former approach is more prone to the risk of overconservativeness and bias in case of small-to-negligible heterogeneity, while the latter to the one of anti-conservativeness and bias in case of high heterogeneity.[16] Overall, an underestimation of the variance, along with the exclusion of small trials (typically associated with a higher heterogeneity[13]), might have contributed to the lower estimated heterogeneity found by Ostuzzi *et al*.

Fourth, the two statistical analyses differ in the effect size measure used. Ostuzzi *et al* analysed risk ratios of relapse, whereas Schneider-Thoma *et al* analysed ORs (and then transformed the NMA results to risk ratios for presentation to increase interpretability). This illustrates that it is currently not clear but vividly discussed which effect size measure should be used in meta-analysis of binary outcomes, since both approaches are prone to criticisms.[17 18]

Fifth, the qualitative assessment of included RCTs and of pooled estimates might suffer from a certain degree of discretion, possibly leading to different data interpretation. The Cochrane Risk of Bias V.2 (RoB2) includes five domains of bias, each of which includes a series of 'signalling questions' to help the researcher judge if the RCT has a 'low' or 'high' risk of bias, or if there are 'some concerns'.[8] After comparing the RoB2 overall judgements of the 44 RCTs included by both working groups for the outcome relapse, we found an agreement of 73.9%, indicating a low inter-rater reliability (k=0.0899, SE=0.1197), consistently with previous observations.[19] Further, in order to assess the overall confidence in pooled estimates of the NMA, the CINeMA (Confidence in Network Meta-Analysis) approach was employed.[20] This methodology is broadly based on the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework, and aims to assess six domains, namely within-study bias (based on the RoB2 judgements), reporting bias, indirectness, imprecision, heterogeneity and incoherence. This approach, although largely automatised through a web-based application, still requires some methodological decisions throughout the process, such as defining the cut-off for clinically relevant effect, and defining how to summarise 'within-study bias' and 'indirectness' across contributions for each network estimate as well as how the overall judgement is reached (because different domains are interconnected and therefore should be considered jointly). Different approaches might often change the overall CINeMA report, affecting certainty in the reliability and applicability of results. Therefore, clinical and policy conclusions strongly based on the certainty of evidence might be criticised as being excessively discretional. As for our example, despite the two NMAs providing largely similar results, conclusion somehow differed. Ostuzzi *et al* indicated three best-performing antipsychotics supported by moderate-to-high certainty of evidence against placebo for both long-acting and oral formulations (table 1), while Schneider-Thoma *et al* refrained from recommending specific antipsychotics because, according to their evaluations, none reached high certainty of evidence, and few statistically clear differences emerged between individual medications.

CONCLUSIONS

NMAs are increasingly carried out in many fields of healthcare, and their results can largely inform the development of clinical practice guidelines and health policy actions. However, such analyses require a number of methodological choices, which can notably vary across different working groups. Although some of these technical issues can be developed further, and hopefully agreed upon by methodologists and be systematically applied to ensure the most accurate results, some other are largely dependent on clinical and research experience, and cannot be easily standardised. This should be regarded as a possibility and not a limit, as long as implications are critically and transparently discussed. It remains imperative that any methodological decisions should be indicated in advance in the study protocol, in order to avoid post-hoc choices based on review findings, and ultimately enhance transparency and replicability of results.

Twitter Giovanni Ostuzzi @psych_verona

Contributors GO, CB, JS-T and SL developed the concept of this paper. GO and JS-T wrote the first draft, which was critically discussed and amended by all authors. FT contributed to writing and revising the statistical and methodological contents. All authors read and approved the final version of the paper.

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 consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Giovanni Ostuzzi http://orcid.org/0000-0003-2248-9524
Johannes Schneider-Thoma http://orcid.org/0000-0002-3448-9532
Stefan Leucht http://orcid.org/0000-0002-4934-4352

REFERENCES

1 Schneider-Thoma J, Chalkou K, Dörries C, *et al*. Comparative efficacy and tolerability of 32 oral and long-acting injectable antipsychotics for the maintenance treatment of adults with schizophrenia: a systematic review and network meta-analysis. *Lancet* 2022;399:824–36.
2 Ostuzzi G, Bertolini F, Tedeschi F, *et al*. Oral and long-acting antipsychotics for relapse prevention in schizophrenia-spectrum disorders: a network meta-analysis of 92 randomized trials including 22,645 participants. *World Psychiatry* 2022;21:295–307.
3 Salanti G, Del Giovane C, Chaimani A, *et al*. Evaluating the quality of evidence from a network meta-analysis. *PLoS ONE* 2014;9:e99682.
4 Moncrieff J, Crellin NE, Long MA, *et al*. Definitions of relapse in trials comparing antipsychotic maintenance with discontinuation or reduction for schizophrenia spectrum disorders: a systematic review. *Schizophrenia Research* 2020;225:47–54.
5 Leucht S, Kane J, Kissling W, *et al*. What does the PANSS mean *Schizophrenia Research* 2005;79:231–8.
6 Czobor P, Sebe B, Acsai K, *et al*. What is the minimum clinically important change in negative symptoms of schizophrenia? PANSS based post-hoc analyses of a phase III clinical trial. *Front Psychiatry* 2022;13:816339.

7 Ostuzzi G, Vita G, Bertolini F, *et al*. Continuing, reducing, switching, or stopping antipsychotics in individuals with schizophrenia-spectrum disorders who are clinically stable: a systematic review and network meta-analysis. *Lancet Psychiatry* 2022;9:614–24.
8 Higgins JP, Thomas J, Chandler J, *et al*. Cochrane Handbook for systematic reviews of interventions version 6.2: Cochrane collaboration; 2021.
9 Higgins JPT, Jackson D, Barrett JK, *et al*. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods* 2012;3:98–110.
10 Papola D, Ostuzzi G, Tedeschi F, *et al*. CBT treatment delivery formats for panic disorder: a systematic review and network meta-analysis of randomised controlled trials. *Psychol Med* 2023;53:614–24.
11 Phillippo D, Ades T, Dias S, *et al*. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE. NICE Decision Support Unit; 2016. Available: https://researchinformation.bris.ac.uk/en/publications/nice-dsu-technical-support-document-18- methods-for-population-adj [Accessed 10 Feb 2023].
12 Efthimiou O, Debray TPA, van Valkenhoef G, *et al*. Getreal in network meta-analysis: a review of the methodology. *Res Synth Methods* 2016;7:236–63.
13 IntHout J, Ioannidis JPA, Borm GF, *et al*. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *J Clin Epidemiol* 2015;68:860–9.
14 Zhang J, Yuan Y, Chu H. The impact of excluding trials from network meta-analyses - an empirical study. *PLoS ONE* 2016;11:e0165889.
15 Turner RM, Bird SM, Higgins JPT. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PLoS One* 2013;8:e59202.
16 Seide SE, Jensen K, Kieser M. A comparison of Bayesian and Frequentist methods in random-effects network meta-analysis of binary data. *Res Synth Methods* 2020;11:363–78.
17 Doi SA, Furuya-Kanamori L, Xu C, *et al*. Controversy and debate: questionable utility of the relative risk in clinical research: Paper 1: a call for change to practice. *J Clin Epidemiol* 2022;142:271–9.
18 Remiro-Azócar A. Purely Prognostic variables may modify marginal treatment effects for non-collapsible effect measures [Arxiv.org]. 2022. Available: https://arxiv.org/abs/2210.01757 [Accessed 10 Feb 2023].
19 Minozzi S, Cinquini M, Gianola S, *et al*. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin Epidemiol* 2020;126:37–44.
20 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.