ELSEVIER

Journal of Clinical Epidemiology 102 (2018) 87–98

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Applying GRADE to a network meta-analysis of antidepressants led to more conservative conclusions

Ashley Bonner[a,*], Paul E. Alexander[a,b], Romina Brignardello-Petersen[a,c], Toshi A. Furukawa[d], Reed A. Siemieniuk[a,e], Yuan Zhang[a], Wojtek Wiercioch[a], Ivan D. Florez[a,f], Yutong Fei[a,g], Arnav Agarwal[a,h], Juan José Yepes-Nuñez[a,i], Joseph Beyene[a], Holger Schünemann[a], Gordon H. Guyatt[a]

[a]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*
[b]*Standards and Practices, Guideline Development, The Infectious Diseases Society of America, Arlington, VA, USA*
[c]*Evidence-Based Dentistry Unit, Faculty of Dentistry, University of Chile, Santiago, Chile*
[d]*Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan*
[e]*Department of Medicine, University of Toronto, Toronto, Canada*
[f]*Department of Pediatrics, University of Antioquia, Medellin, Colombia*
[g]*Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China*
[h]*School of Medicine, University of Toronto, Toronto, Ontario, Canada*
[i]*School of Medicine, University of Antioquia, Medellín, Colombia*

Accepted 3 May 2018; Published online 22 May 2018

Abstract

Objective: To explore the impact of applying the Grading of Recommendations and Assessment, Development, and Evaluation (GRADE) approach to assess the certainty of the evidence in a published network meta-analysis (NMA) of antidepressant therapies.

Study design and settings: We applied the GRADE approach to rate the certainty of the evidence for two outcomes, efficacy and acceptability, in each of the 66 paired comparisons within a previously published NMA assessing the relative efficacy and acceptability of 12 new-generation antidepressants.

Results: For the outcome of efficacy, of the 25 comparisons in which the 95% CrI of OR excluded 1, 18 had certainty of evidence rated high or moderate. For the outcome of acceptability, of the 13 comparisons whose 95% CrI excluded 1, 10 had certainty of evidence rated high or moderate. Of the 11 comparisons involving sertraline, the antidepressants that the authors of the NMA suggested to be best, only 3 demonstrated it to be more effective and only 3 showed better tolerance, based on a 95% CrI excluding 1 and a high or moderate rating of certainty.

Conclusions: In this example, application of GRADE highlighted varying evidence certainty, led to more conservative conclusions, and potentially avoided unwarranted strong inferences based on low certainty evidence.

Keywords: Certainty of the evidence; GRADE approach; Network meta-analysis; Results interpretation

1. Introduction

Optimal clinical decision-making requires systematic summaries of the best available evidence [1]. Systematic reviews with meta-analysis of randomized clinical trials (RCTs) represent the most compelling and trustworthy of such summaries. Traditional pairwise meta-analysis provides an estimate of relative effect of two interventions by pooling the results from multiple RCTs that have compared those interventions [2–4]. When clinicians must choose between multiple treatment options for the same patient—as it is often the case—traditional pairwise meta-analysis provides limited guidance.

Network meta-analysis (NMA), an extension of the traditional pairwise meta-analysis, provides simultaneous comparisons of 3 or more interventions. NMA combines direct and indirect evidence across multiple trials that compare any of the interventions included in the network. NMA provides relative estimates of effect for interventions that have never been directly compared and may result in improved

* Corresponding author. 1280 Main Street W, Hamilton, ON L8S 4L8, Canada. Tel.:+1 9055259140 x26854
E-mail address: ashleybonner8@gmail.com (A. Bonner).

https://doi.org/10.1016/j.jclinepi.2018.05.009
0895-4356/

What is New?

Key findings

- In this example, using the Grading of Recommendations and Assessment, Development, and Evaluation (GRADE) approach to assess the certainty of the evidence from a network meta-analysis (NMA) led to more conservative inferences regarding the relative merits of the interventions under consideration.

What this adds to what was known?

- To date, most authors of NMAs have considered only the relative estimates of effect and rankings to make conclusions about which treatments are more effective than others, and have failed to address the certainty in the evidence when drawing their conclusions.
- This example illuminates the importance of assessing the certainty of evidence from a NMA.

What is the implication and what should change now?

- Using the GRADE approach to assess the certainty of the evidence from NMA may result in important differences in the conclusions from an NMA; using GRADE is likely to avoid misleadingly strong inferences regarding interventions' relative benefit and harm.
- NMA users are likely to be better informed when authors use GRADE principles to rate and report certainty in the relative effects from their NMA.

precision through combining direct and indirect estimates. A potential one-stop shop to inform decisions, some have touted NMA as a revolutionary tool for health technology assessments [5,6], and its use is rapidly increasing [7].

NMA remains a new methodology, and practice continues to evolve. To date, only a few NMAs have presented the certainty of evidence for each paired comparison [8−11]. This limitation is important: some pairwise comparisons are much more trustworthy than others, and the relative trustworthiness should guide patient management decisions.

The Grading of Recommendations and Assessment, Development, and Evaluation (GRADE) working group has developed a widely used framework to address the certainty of evidence for patient important outcomes from systematic reviews and meta-analyses [12−14]. This framework has been extended to NMAs [15], addressing additional challenges including the particular vulnerabilities of indirect comparison estimates (intransitivity) and incoherence between direct and indirect evidence [16].

In 2009, Cipriani et al. [17] published a systematic review and NMA to evaluate the relative efficacy and acceptability of 12 new-generation antidepressants in the acute-phase (8 weeks) treatment of adults with unipolar major depressive disorder. This systematic review was rigorous in its application of standard systematic review methodology and in dealing with the statistical challenges of NMA. It generated considerable attention and some controversy [18−23]. Published 2 years before the present standard Cochrane risk of bias tool was officially announced [24,25], and 5 years before any guideline for the application of GRADE framework to NMA was available [15,26], it did not offer a rating of evidence certainty for the paired comparisons that comprised the NMA. For all these reasons, and because an author of the original NMA (TAF) expressed enthusiasm for participating in a reassessment, we chose this NMA to explore the potential impact of the GRADE approach to rate the certainty in evidence on inferences arising from an NMA.

The sole purpose of this investigation is to explore the impact of applying GRADE ratings of certainty to an existing NMA. There may be much new evidence available in the 8 years since its publication. Readers should not, therefore, use the results to guide their practice with respect to antidepressant use, but rather consider the implications of our application of GRADE to this previous NMA.

2. Methods

2.1. NMA comparing 12 new-generation antidepressants

Cipriani et al. [17] describe their methods in detail in their publication; here, we summarize briefly. The authors included RCTs comparing bupropion, citalopram, duloxetine, escitalopram, fluoxetine, fluvoxamine, milnacipran, mirtazapine, paroxetine, reboxetine, sertraline, and venlafaxine against any of the others. RCTs deemed inadequate with regard to random allocation concealment, and blinding were excluded. The authors identified 117 eligible studies, of which 111 contributed data for an analysis of efficacy and 112 for an analysis of acceptability.

The authors defined efficacy as the proportion of patients who achieved a reduction of at least 50% on the Hamilton depression rating scale or, if unavailable, as the proportion of patients who achieved the same reduction on the Montgomery-Asberg Depression Rating Scale or who scored 'much improved' or 'very much improved' on the Clinical Global Impression scale. The primary outcome of acceptability was defined as the proportion of patients who discontinued their participation in the study for any reason. Of the 66 possible pairwise comparisons for each outcome, 42 had direct evidence for efficacy, and 41 had direct evidence for acceptability (Appendix Figures 1 and 2). The authors of the systematic review looked at both outcomes around 8 weeks (range 6-12 weeks) after the start of the treatment.

Cipriani et al. [17] concluded that mirtazapine, escitalopram, venlafaxine, and sertraline were more efficacious than duloxetine, fluoxetine, fluvoxamine, and paroxetine and that reboxetine was less efficacious that all the other drugs. They also concluded that escitalopram and sertraline were better in terms of acceptability than duloxetine, fluvoxamine, paroxetine, reboxetine, and venlafaxine and that, overall, escitalopram and sertraline were the best treatments.

2.2. Study selection and data abstraction

Because our aim was to explore the possible impact of applying the GRADE approach to rate the certainty of the evidence from the NMA, we used the same studies, treatments, and outcomes included in the original NMA [17]. In addition to using the data published by the authors, we abstracted study level data that were necessary to apply the GRADE certainty ratings. For each trial, we captured information on publication status and sponsorship, as well as possible effect modifiers—dose, specific age restrictions, and whether or not enrollment was restricted to patients with anxiety. To facilitate our cross-treatment comparisons, the original NMA authors provided dose levels, both as ranges used and as classifications [17,27–29] of 'low,' 'medium,' and 'high'.

We developed a data abstraction form that included explicit definitions of each trial characteristic. Two independent reviewers abstracted data for the first 20 trials to calibrate the process and ensure agreement. After independent abstraction, discrepancies were discussed and resolved (3rd party adjudication was available if necessary); consensus was reached for 100% of the items. One reviewer who had participated in calibration (AB) conducted the remaining data abstraction.

2.3. Data analysis

We used the same trial-level data and the analytic approach to NMA as the original authors [17]. We performed NMA using a Bayesian approach to calculate pairwise comparison estimates and their corresponding 95% credible intervals using a random-effects model [30,31]. In addition, for each pairwise comparison, we obtained direct and indirect estimates of effect using a node-splitting approach [31,32]. We used the package *gemtc* [33] in R for all analyses [34].

2.4. Application of the GRADE approach for network meta-analysis

For each outcome and pairwise comparison, we used the GRADE approach to rate the certainty of the evidence from the NMA [15]. Appendix Figure 3 depicts the process we used to rate certainty in the network estimates. The GRADE process specific for this NMA (criteria, thresholds, and so forth) was developed and recorded prior to execution. For each stage of the process, five teams of two reviewers (A.B. and A.A.; R.B. and W.W.; I.F. and J.Y.; R.S. and Y.Z.; and P.A. and Y.F.) conducted the GRADE assessments. After two instructional meetings involving all members, each team of two was assigned a set of pairwise comparisons and carried out GRADE evaluations independently and in duplicate, resolving discrepancies via discussion or third-party assistance (A.B., R.B., and G.G.).

2.4.1. Rating the certainty of the direct evidence to inform the network estimate rating

We used GRADE's well-established approach to rating certainty in RCTs from direct comparisons [12,35,36]. To inform the network estimate rating, when rating the direct evidence, we evaluated the risk of bias [37], inconsistency [38], indirectness [39], and publication bias [40]. Because the credible interval of the network estimate may differ from that of the direct estimate, we did not judge imprecision at this stage [16].

For each study, two reviewers independently evaluated risk of bias using a modified version of the Cochrane risk of bias instrument [41]. We considered the domains of sequence generation, allocation concealment, blinding of participants and personnel, and incomplete outcome data. For the latter, we judged a study as high risk of bias for the efficacy outcome due to incomplete outcome data if more than 10% of enrolled participants were lost to follow-up. However, we did not judge such studies as high risk of bias due to incomplete data with regard to the acceptability outcome because acceptability involved counting patients who were lost to follow-up. We judged a study as high risk of bias if any of the domains were judged as high risk of bias.

We rated down a direct estimate in the following cases: (1) risk of bias: one or more studies were at high risk of bias and (a) estimates from high and low risk of bias studies agree (same conclusions regarding relative effectiveness of treatments) but low risk of bias studies contribute less than 50% of the weight to the pooled estimate (i.e., insufficient evidence to render the low risk of bias studies sufficiently credible) or (b) estimates from high and low risk of bias studies disagree (different conclusions regarding relative effectiveness of treatments) and low risk of bias studies contribute less than 75% of the weight to the pooled estimate, (2) inconsistency: visual or statistical methods suggested important between-study inconsistency, (3) indirectness: more than 30% of the weight of the pooled estimate came from studies in which there was a large difference in dose level ('low' vs. 'high' or 'low' vs. 'moderate-high') between the treatments, raising concerns about the applicability of the estimate, and (4) publication bias: all studies were published and more than 70% of the weight of the pooled estimate comes from studies funded by the manufacturer company for which the pooled estimate shows favorable evidence.

2.4.2. Rating the certainty of the indirect evidence to inform the network rating

We rated the certainty of the indirect evidence using the most dominant first order loop that contributed to the network estimate. For each pairwise comparison, we

identified all the first order loops (i.e., all pathways that compared the two treatments of interest via a common comparator). When there were 2 or more first order loops, we identified the one that contributed the most to the network estimate using Bucher's method [42] to calculate the variance of each indirect estimate. We then used the loop with the smallest variance.

We obtained the rating of the indirect evidence using the lowest of the ratings (without considering imprecision) [16] of the two direct comparisons from the loop contributing to the indirect estimate and evaluated intransitivity. We rated down for intransitivity when important imbalances in dose, age, and presence of anxiety occurred between the two direct comparisons. This threshold was more stringent [15] than when we rated direct paired comparisons, in which we did not rate down when studies varied in dose, age, and presence of anxiety.

2.4.3. Rating the certainty of the network estimate

The evidence certainty rating of whatever comparison source (direct or indirect) had the higher certainty rating [15] represented the initial network estimate certainty rating which could then be modified (rated down) by considering imprecision. We did not modify (rate down) the certainty rating of the network estimate twice when both intransitivity and incoherence were present because incoherence can be considered the statistical manifestation of intransitivity.

We evaluated incoherence between the direct and indirect estimates considering the extent of difference between the point estimates, the extent of overlap between the confidence intervals, and the statistical test of incoherence using the node-splitting approach [43] with P-value < 0.10.

We evaluated imprecision using the network estimate; if the 95% credible interval excluded an odds ratio (OR) of 1, we did not rate down for imprecision. When the results did not exclude an OR of 1, we rated down for imprecision if the lower boundary of the credible interval was below 0.8 or the upper boundary was above 1.25.

2.5. Drawing conclusions using the results from the NMA and the GRADE certainty ratings

Network estimates, being more precise than direct and indirect, tend to have higher certainty, but when incoherence is present, the direct or indirect estimate may have higher certainty than the network estimate. GRADE guidance suggests two options for dealing with such situations. One may choose the highest certainty estimate and use that certainty rating or the network estimate. In this case, to align our methods most closely to the original publication, we chose the latter. To infer superiority with respect to an outcome (efficacy: "more effective"; acceptability: "better tolerated"), we required the same condition as the original authors: the relative treatment effect credible interval to exclude an OR of 1. In addition, however, we required high or moderate certainty in the evidence for that pairwise comparison.

3. Results

We included 117 trials, of which 111 contributed data for an analysis of efficacy and 112 contributed data for analysis of acceptability. Appendix Figures 1 and 2 present the geometry of the network for each outcome. Table A reports the risk of bias assessments for pairwise comparisons with direct evidence.

3.1. Certainty of the evidence for efficacy

Table 1 displays certainty ratings with network estimates and credible intervals for all comparisons with the efficacy outcome, alongside final (precision included at the end) certainty ratings for contributing direct and indirect estimates. Table 1A in the appendix shows certainty ratings for direct and indirect ratings as they were used to inform the network estimate certainty rating (precision excluded; as described in our Methods Section).

From the 66 network estimates, after completing assessments for all criteria, 4 (6.1%) proved of high certainty, 28 (42.4%) moderate certainty, 15 (22.7%) low certainty, and 19 (28.8%) very low certainty (Table 1). Twenty-seven (40.9%) of the comparisons had been rated down for issues with intransitivity and 38 (57.6%) for imprecision. Four (9.5%) of the 42 comparisons with both direct and indirect evidence showed incoherence.

Of the 42 comparisons with both direct and indirect evidence, 3 (7.1%) had a network estimate with less certainty than that of the direct estimate (precision reincluded), 22 (52.4%) had a network estimate and direct estimate with equal certainty, and 17 (40.5%) had a network estimate with higher certainty than that of the direct estimate (Table 1). Of the four comparisons displaying incoherence, two had a direct estimate with much higher certainty than the indirect estimate, leading us to suggest that, for these comparisons, the direct estimate is more trustworthy than its indirect and network counterparts. These comparisons were citalopram vs. escitalopram (direct: high certainty and indirect: very low certainty) and mirtazapine vs. venlafaxine (direct: moderate certainty and indirect: very low certainty).

3.2. Certainty of the evidence for acceptability

Table 2 displays certainty ratings with network estimates and credible intervals for all comparisons with the acceptability outcome, alongside final (precision included at the end) certainty ratings for contributing direct and indirect estimates. Table 2A in the appendix shows certainty ratings for direct and indirect ratings as they were used to inform the network estimate certainty rating (precision excluded; as described in our Methods Section).

From the 66 network estimates, our assessments resulted in 10 (15.2%) with high certainty, 26 (39.4%) with moderate certainty, 21 (31.8%) with low certainty, and 9 (13.6%) with very low certainty (Table 2). Twenty-three (34.8%) of

Table 1. Direct, Indirect, and Network Estimates with GRADE ratings for all 66 pairwise comparisons within the network of antidepressants for the outcome Efficacy

Treatment 1	Treatment 2	Direct Est.	Certainty	Indirect Est.	Certainty	Network Est.	Certainty
Fluoxetine	Bupropion	0.82 (0.62, 1.10)	Low 1,p	1.21 (0.98, 1.49)	Very low (1,4)p	1.08 (0.91, 1.30)	Very low (D:1) p,6
Fluoxetine	Citalopram	1.05 (0.77, 1.43)	Low 1,p	1.12 (0.92, 1.37)	Low 5,p	1.10 (0.93, 1.31)	Low (D:1/I:5) p
Fluoxetine	Duloxetine	1.01 (0.44, 2.32)	Very low 3,4,p	0.99 (0.78, 1.25)	Very low (2,3) 5,p	0.99 (0.79, 1.24)	Very low (D:3,4) p
Fluoxetine	Escitalopram	1.23 (0.87, 1.74)	Moderate p	1.34 (1.12, 1.60)	Moderate (1)	1.32 (1.13, 1.55)	High (D:)
Fluoxetine	Fluvoxamine	1.03 (0.64, 1.66)	Moderate p	0.97 (0.74, 1.27)	Low (4/1)p	0.98 (0.77, 1.24)	Moderate (D:) p
Fluoxetine	Milnacipran	0.87 (0.53, 1.41)	Very low 2,3,p	1.21 (0.78, 1.89)	Low (1/1)p	1.01 (0.76, 1.35)	Low (I:(1/1)) p
Fluoxetine	Mirtazapine	1.51 (1.09, 2.11)	Moderate 4	1.30 (1.04, 1.65)	Moderate (1/4)	1.38 (1.14, 1.67)	Moderate (D:4/I:(1/4))
Fluoxetine	Paroxetine	1.02 (0.82, 1.26)	Low 1,p	1.06 (0.88, 1.28)	Low (4/4)p	1.01 (0.89, 1.17)	Moderate (D:1/I:(4/4))
Fluoxetine	Reboxetine	0.72 (0.52, 1.01)	Low 1,p	0.61 (0.41, 0.91)	Low (1/4)5	0.67 (0.53, 0.87)	Moderate (D:1)
Fluoxetine	Sertraline	1.36 (1.06, 1.74)	High	1.15 (0.95, 1.40)	Very low (1,4)p	1.25 (1.07, 1.46)	High (D:)
Fluoxetine	Venlafaxine	1.36 (1.14, 1.62)	Low 1,4	1.19 (0.98, 1.46)	Low (1)p	1.28 (1.11, 1.47)	Moderate (I:(1))
Bupropion	Citalopram	NA	NA	1.02 (0.81, 1.28)	Very low (1/1) 5,p	1.02 (0.81, 1.28)	Very low (I:(1/1)5) p
Bupropion	Duloxetine	NA	NA	0.92 (0.70, 1.20)	Very low (2/2) 5,p	0.92 (0.70, 1.19)	Very low (I:(2/2)5) p
Bupropion	Escitalopram	1.07 (0.69, 1.67)	Low 2,p	1.29 (1.01, 1.64)	Low (1)5	1.22 (1.00, 1.50)	Low (D:2) p
Bupropion	Fluvoxamine	NA	NA	0.91 (0.68, 1.20)	Very low (1)5,p	0.91 (0.69, 1.20)	Very low (I:(1)5) p
Bupropion	Milnacipran	NA	NA	0.93 (0.67, 1.30)	Very low (2,3) 5,p	0.93 (0.66, 1.30)	Very low (I:(2,3)5) p
Bupropion	Mirtazapine	NA	NA	1.28 (1.00, 1.63)	Very low (1/4) 5,p	1.27 (1.00, 1.63)	Very low (I:(1/4)5) p
Bupropion	Paroxetine	1.37 (0.56, 3.36)	Moderate p	0.93 (0.75, 1.13)	Low (1/1)p	0.94 (0.77, 1.15)	Moderate (D:) p
Bupropion	Reboxetine	NA	NA	0.63 (0.47, 0.84)	Low (1/1)5	0.62 (0.46, 0.84)	Low (I:(1/1)5)
Bupropion	Sertraline	0.93 (0.69, 1.27)	Low 3,p	1.28 (1.01, 1.63)	Moderate (1)	1.15 (0.96, 1.40)	Low (D:3/I:(1)) p
Bupropion	Venlafaxine	1.17 (0.87, 1.59)	Moderate p	1.19 (0.93, 1.52)	Very low (1,4)p	1.18 (0.99, 1.42)	Moderate (D:) p
Citalopram	Duloxetine	NA	NA	0.90 (0.69, 1.16)	Low (2)p	0.90 (0.69, 1.16)	Low (I:(2)) p
Citalopram	Escitalopram	1.48 (1.16, 1.89)	High	0.96 (0.75, 1.23)	Very low (1)5,p	1.20 (1.00, 1.43)	Low (D:) p,6
Citalopram	Fluvoxamine	0.90 (0.50, 1.62)	Moderate p	0.89 (0.66, 1.19)	Low (1)p	0.89 (0.68, 1.16)	Moderate (D:) p
Citalopram	Milnacipran	NA	NA	0.91 (0.66, 1.27)	Very low (2,3)p	0.91 (0.66, 1.27)	Very low (I:(2,3)) p
Citalopram	Mirtazapine	0.76 (0.38, 1.52)	Moderate p	1.32 (1.03, 1.70)	Moderate (1/4)	1.25 (0.99, 1.58)	Moderate (D:) p
Citalopram	Paroxetine	0.65 (0.44, 0.96)	Moderate 4	0.99 (0.81, 1.21)	Moderate (1/1)	0.92 (0.76, 1.11)	Very low (D:4/I:(1/1)) p,6
Citalopram	Reboxetine	0.58 (0.34, 0.99)	Moderate 4	0.65 (0.46, 0.92)	Moderate (1/1)	0.61 (0.47, 0.80)	Moderate (D:4/I:(1/1))
Citalopram	Sertraline	1.07 (0.70, 1.64)	Moderate p	1.14 (0.92, 1.43)	Low (1)p	1.13 (0.93, 1.39)	Moderate (D:)

(*Continued*)

Table 1. Continued

Treatment 1	Treatment 2	Direct Est.	Certainty	Indirect Est.	Certainty	Network Est.	Certainty
							p
Citalopram	Venlafaxine	0.91 (0.46, 1.78)	Moderate p	1.18 (0.97, 1.46)	Very low (1,4)p	1.16 (0.95, 1.42)	Moderate (D:) p
Duloxetine	Escitalopram	1.30 (0.89, 1.90)	Low 2,p	1.45 (1.04, 2.00)	Very low (2,3/ 1,2)5	1.33 (1.07, 1.66)	Moderate (D:2)
Duloxetine	Fluvoxamine	NA	NA	1.00 (0.73, 1.35)	Very low (2,3) 5,p	0.99 (0.73, 1.35)	Very low (I:(2,3)5) p
Duloxetine	Milnacipran	NA	NA	1.02 (0.71, 1.45)	Very low (2,3)p	1.02 (0.71, 1.46)	Very low (I:(2,3)) p
Duloxetine	Mirtazapine	NA	NA	1.39 (1.06, 1.84)	Very low (2,3)5	1.39 (1.06, 1.85)	Very low (I:(2,3)5)
Duloxetine	Paroxetine	1.10 (0.74, 1.63)	Very low 2,3,p	0.95 (0.70, 1.30)	Very low (1,2) 5,p	1.02 (0.83, 1.28)	Very low (D:2,3) p
Duloxetine	Reboxetine	NA	NA	0.68 (0.49, 0.94)	Low (3,4)	0.68 (0.49, 0.95)	Low (I:(3,4))
Duloxetine	Sertraline	NA	NA	1.26 (0.98, 1.62)	Very low (1,3) 5,p	1.26 (0.98, 1.62)	Very low (I:(1,3)5) p
Duloxetine	Venlafaxine	NA	NA	1.29 (1.01, 1.65)	Very low (2,3)5	1.29 (1.01, 1.66)	Very low (I:(2,3)5)
Escitalopram	Fluvoxamine	NA	NA	0.75 (0.57, 0.97)	Moderate 5	0.74 (0.57, 0.96)	Moderate (I:5)
Escitalopram	Milnacipran	NA	NA	0.76 (0.55, 1.05)	Very low (1,2) 5,p	0.76 (0.55, 1.05)	Very low (I:(1,2)5) p
Escitalopram	Mirtazapine	NA	NA	1.04 (0.83, 1.32)	Very low (4)5,p	1.05 (0.83, 1.31)	Very low (I:(4)5) p
Escitalopram	Paroxetine	0.89 (0.61, 1.32)	Very low 1,2,p	0.74 (0.61, 0.89)	Moderate (1)	0.77 (0.65, 0.91)	Moderate (I:(1))
Escitalopram	Reboxetine	NA	NA	0.51 (0.39, 0.67)	Low (1)5	0.51 (0.39, 0.67)	Low (I:(1)5)
Escitalopram	Sertraline	1.12 (0.77, 1.61)	Very low 1,3,p	0.91 (0.74, 1.11)	Moderate p	0.95 (0.79, 1.14)	Moderate (I:) p
Escitalopram	Venlafaxine	0.83 (0.47, 1.44)	Moderate p	1.00 (0.82, 1.22)	Very low (1,4)5	0.97 (0.81, 1.16)	High (D:)
Fluvoxamine	Milnacipran	1.76 (0.81, 3.83)	Low 1,p	0.91 (0.62, 1.32)	Low (1)p	1.03 (0.72, 1.45)	Low (D:1/I:(1)) p
Fluvoxamine	Mirtazapine	1.14 (0.76, 1.70)	Low 1,p	1.55 (1.12, 2.13)	Moderate (4)	1.41 (1.08, 1.84)	Moderate (D:1/I:(4))
Fluvoxamine	Paroxetine	1.21 (0.74, 1.96)	Moderate p	0.99 (0.75, 1.32)	Low (1/4)p	1.04 (0.82, 1.32)	Moderate (D:) p
Fluvoxamine	Reboxetine	NA	NA	0.69 (0.49, 0.96)	Moderate (1)	0.69 (0.49, 0.95)	Moderate (I:(1))
Fluvoxamine	Sertraline	0.83 (0.36, 1.88)	Very low 1,2,p	1.39 (1.05, 1.83)	High	1.27 (0.99, 1.64)	Moderate (I:) p
Fluvoxamine	Venlafaxine	2.36 (1.04, 5.39)	Low 1,3	1.22 (0.94, 1.60)	Very low (1,4)p	1.30 (1.01, 1.68)	Low (D:1,3/I:(1,4))
Milnacipran	Mirtazapine	NA	NA	1.37 (0.97, 1.91)	Low (1/4)p	1.37 (0.98, 1.90)	Low (I:(1/4)) p
Milnacipran	Paroxetine	1.05 (0.67, 1.66)	Low 1,p	0.99 (0.69, 1.43)	Very low (2,3) 5,p	1.01 (0.75, 1.36)	Low (D:1) p
Milnacipran	Reboxetine	NA	NA	0.67 (0.46, 0.97)	Low (2,3)	0.67 (0.46, 0.97)	Low (I:(2,3))
Milnacipran	Sertraline	0.48 (0.08, 2.87)	Low 1,p	1.28 (0.94, 1.78)	Very low (2,3) 5,p	1.24 (0.91, 1.72)	Low (D:1) p
Milnacipran	Venlafaxine	NA	NA	1.27 (0.92, 1.73)	Very low (2,3/ 1,4)5,p	1.27 (0.93, 1.74)	Very low (I:(2,3/1,4) 5) p
Mirtazapine	Paroxetine	0.79 (0.59, 1.06)	Low 4,p	0.71 (0.55, 0.91)	Moderate (4/1)	0.74 (0.60, 0.90)	Moderate (D:4/I:(4/1))
Mirtazapine	Reboxetine	NA	NA	0.49 (0.36, 0.66)	Moderate (4/1)	0.49 (0.36, 0.67)	Moderate (I:(4/1))
Mirtazapine	Sertraline	1.03 (0.66, 1.61)	Moderate p	0.88 (0.69, 1.12)	Low (4)p	0.91 (0.73, 1.13)	Moderate (D:) p

(*Continued*)

Table 1. Continued

Treatment 1	Treatment 2	Direct Est.	Certainty	Indirect Est.	Certainty	Network Est.	Certainty
Mirtazapine	Venlafaxine	0.66 (0.44, 0.97)	Moderate 4	1.04 (0.82, 1.31)	Very low (1,4) 5,p	0.93 (0.75, 1.14)	Very low (D:4) p,6
Paroxetine	Reboxetine	NA	NA	0.66 (0.50, 0.86)	Moderate (1/1)	0.66 (0.50, 0.87)	Moderate (I:(1/1))
Paroxetine	Sertraline	1.83 (0.90, 3.71)	Low 2,p	1.16 (0.95, 1.40)	Low (1)p	1.23 (1.03, 1.46)	Moderate (D:2/I:(1))
Paroxetine	Venlafaxine	1.12 (0.74, 1.71)	Moderate p	1.28 (1.07, 1.52)	Low (1,4)	1.26 (1.06, 1.49)	High (D:)
Reboxetine	Sertraline	1.37 (0.41, 4.54)	Very low 1,3,p	1.88 (1.41, 2.51)	Moderate (1)	1.86 (1.40, 2.46)	Moderate (I:(1))
Reboxetine	Venlafaxine	2.22 (0.98, 5.05)	Low 1,p	1.86 (1.40, 2.49)	Very low (1,4)5	1.90 (1.44, 2.48)	Moderate (D:1)
Sertraline	Venlafaxine	1.16 (0.83, 1.62)	Low 1,p	0.99 (0.81, 1.20)	Low (1,4)	1.02 (0.86, 1.21)	Moderate (D:1)

NA = Not applicable; there were no studies directly comparing the two treatments.
1 = Rated down due to issues with risk of bias.
2 = Rated down due to issues with consistency.
3 = Rated down due to issues with directness.
4 = Rated down due to issues with publication bias.
5 = Rated down due to issues with transitivity.
In this table, the direct and indirect certainty ratings have the precision domain reincluded for fair comparison with network certainty ratings. In contrast, Appendix Table 1A shows certainty ratings for direct and indirect estimates that were used to inform (precision excluded) the certainty of the network estimate
p = Rated down due to issues with precision; note that any rating down due to precision was done as a final step for direct, indirect, and network estimates.
6 = Rated down due to issues with coherence; 6* = evidence of incoherence but we did not rate down because the network estimate certainty was informed by an indirect estimate which had already been rated down for intransitivity.
(#) = Contributing evidence was rated down due to issues with #, which affects the rating for the current estimate.
D: = Network estimate rating initiated by rating of direct estimate.
I: = Network estimate rating initiated by rating of indirect estimate.
#/# = Used to show down-ratings when two contributing evidence sources had the same certainty rating and therefore both governed the initial rating for the current estimate.
Example in Indirect Est. column: (3,4/1,4) means both pairwise comparisons in the most dominant first order loop had the same certainty rating; the first comparison was rated down due to issues with directness and publication bias, and the second comparison was rated down for issues with risk of bias and publication bias.
Example in Network Est. column: (D:1,2/I:(1)5) means both direct estimate and indirect estimate had the same certainty rating; the direct estimate was rated down for issues with risk of bias and consistency, and the indirect estimate was rated down for issues with risk of bias and transitivity.

the comparisons had been rated down for issues with intransitivity and 52 (78.8%) for imprecision. Five (12.2%) of the 41 comparisons with both direct and indirect evidence showed incoherence.

Of the 41 comparisons with both direct and indirect evidence, 4 (9.8%) had a network estimate with less certainty than the direct estimate, 24 (58.5%) had a network estimate and direct estimate with equal certainty, and 13 (31.7%) had a network estimate with higher certainty than the direct estimate. None of the five comparisons displaying incoherence showed a large enough difference in certainty rating between direct and indirect estimates to suggest trusting the direct or indirect estimate over the network estimate.

3.3. Pairwise comparison in which one treatment was superior than the other

For efficacy, Cipriani et al. [17] summarized that mirtazapine, escitalopram, venlafaxine, and sertraline were more effective than duloxetine, fluoxetine, fluvoxamine, and paroxetine and that reboxetine was less effective than all the other drugs. Based on their analysis, one treatment was more effective than another in 25 (37.9%) of the 66 pairwise comparisons.

For efficacy, using our criteria, we found that 18 (27.3%) of the 66 pairwise comparisons showed one treatment as being more effective than another. Mirtazapine was more effective than fluoxetine, fluvoxamine, paroxetine; escitalopram was more effective than duloxetine, fluoxetine, fluvoxamine, and paroxetine; venlafaxine was more effective than fluoxetine and paroxetine; sertraline was more effective than fluoxetine and paroxetine; and reboxetine was less effective than citalopram, fluoxetine, fluvoxamine, mirtazapine, paroxetine, sertraline, and venlafaxine. Owing to considering the certainty in the evidence, our conclusion about the relative effectiveness of one treatment against another differed in 7 (28.0%) of 25 comparisons declared significant by the original authors.

For acceptability, Cipriani et al. [17] summarized that escitalopram and sertraline were better in terms of acceptability than duloxetine, fluvoxamine, paroxetine, reboxetine, and

Table 2. Direct, indirect, and network estimates with GRADE certainty ratings for all 66 pairwise comparisons within the network of antidepressants for the outcome acceptability

Treatment 1	Treatment 2	Direct Est.	Certainty	Indirect Est.	Certainty	Network Est.	Certainty
Fluoxetine	Bupropion	1.01 (0.75, 1.36)	Moderate p	0.85 (0.67, 1.07)	Moderate p	0.90 (0.74, 1.08)	Moderate (D:/I:) p
Fluoxetine	Citalopram	1.17 (0.80, 1.70)	Moderate p	0.83 (0.66, 1.05)	Moderate p	0.90 (0.73, 1.09)	Moderate (D:/I:) p
Fluoxetine	Duloxetine	0.91 (0.38, 2.17)	Low 3,p	1.22 (0.92, 1.63)	Very low (3)5,p	1.20 (0.90, 1.57)	Low (D:3) p
Fluoxetine	Escitalopram	0.98 (0.37, 2.56)	Low 2,p	0.82 (0.67, 1.01)	Moderate p	0.84 (0.70, 1.01)	Moderate (I:) p
Fluoxetine	Fluvoxamine	1.17 (0.66, 2.09)	Moderate p	1.24 (0.92, 1.70)	Moderate p	1.22 (0.93, 1.62)	Moderate (D:/I:) p
Fluoxetine	Milnacipran	1.02 (0.71, 1.46)	Low 3,p	1.06 (0.64, 1.71)	Moderate p	1.03 (0.76, 1.41)	Moderate (I:) p
Fluoxetine	Mirtazapine	1.05 (0.57, 1.95)	Moderate p	1.02 (0.79, 1.32)	Low (4)p	1.04 (0.82, 1.30)	Moderate (D:) p
Fluoxetine	Paroxetine	1.05 (0.89, 1.23)	High	1.22 (0.98, 1.52)	Moderate p	1.11 (0.95, 1.27)	Moderate (D:/I:) p
Fluoxetine	Reboxetine	1.47 (1.07, 2.02)	High	1.34 (0.84, 2.11)	Low (1)p	1.43 (1.08, 1.89)	High (D:)
Fluoxetine	Sertraline	0.81 (0.58, 1.12)	Moderate p	0.92 (0.73, 1.15)	Moderate p	0.89 (0.73, 1.06)	Moderate (D:/I:) p
Fluoxetine	Venlafaxine	1.07 (0.88, 1.29)	Moderate p	1.07 (0.85, 1.34)	Moderate p	1.07 (0.92, 1.24)	High (D:/I:)
Bupropion	Citalopram	NA	NA	1.00 (0.78, 1.29)	Moderate p	1.00 (0.78, 1.29)	Moderate (I:) p
Bupropion	Duloxetine	NA	NA	1.34 (0.98, 1.82)	Very low (3)5,p	1.34 (0.98, 1.82)	Very low (I:(3)5) p
Bupropion	Escitalopram	1.02 (0.75, 1.39)	Moderate p	0.89 (0.68, 1.18)	Moderate p	0.94 (0.75, 1.17)	Moderate (D:/I:) p
Bupropion	Fluvoxamine	NA	NA	1.37 (1.00, 1.88)	Low 5,p	1.37 (1.00, 1.88)	Low (I:5) p
Bupropion	Milnacipran	NA	NA	1.15 (0.81, 1.65)	Very low (3)5,p	1.15 (0.81, 1.65)	Very low (I:(3)5) p
Bupropion	Mirtazapine	NA	NA	1.15 (0.88, 1.52)	Low (4)p	1.15 (0.88, 1.52)	Low (I:(4)) p
Bupropion	Paroxetine	1.16 (0.62, 2.20)	Moderate p	1.25 (0.99, 1.56)	Moderate p	1.23 (1.00, 1.53)	Moderate (D:/I:) p
Bupropion	Reboxetine	NA	NA	1.60 (1.15, 2.22)	Moderate 5	1.60 (1.15, 2.22)	Moderate (I:5)
Bupropion	Sertraline	1.51 (0.86, 2.64)	Low 4,p	0.85 (0.66, 1.09)	Moderate p	0.99 (0.79, 1.25)	Low (I:) p,6
Bupropion	Venlafaxine	1.01 (0.76, 1.32)	Moderate p	1.34 (1.04, 1.74)	High	1.19 (0.97, 1.46)	Moderate (D:/I:) p
Citalopram	Duloxetine	NA	NA	1.34 (0.98, 1.82)	Very low (3)5,p	1.34 (0.98, 1.82)	Very low (I:(3)5) p
Citalopram	Escitalopram	0.86 (0.64, 1.17)	Moderate p	1.03 (0.76, 1.38)	Very low (3)5,p	0.94 (0.76, 1.16)	Moderate (D:) p
Citalopram	Fluvoxamine	1.42 (0.75, 2.67)	Moderate p	1.35 (0.97, 1.91)	Moderate p	1.37 (1.01, 1.86)	High (D:/I:)
Citalopram	Milnacipran	NA	NA	1.15 (0.81, 1.64)	Low (3) p	1.15 (0.81, 1.64)	Low (I:(3)) p
Citalopram	Mirtazapine	2.36 (0.99, 5.64)	Low 4,p	1.07 (0.80, 1.42)	Very low (4)5,p	1.16 (0.88, 1.52)	Very low (D:4) p,6
Citalopram	Paroxetine	0.99 (0.61, 1.60)	Moderate p	1.28 (1.01, 1.62)	High	1.24 (1.00, 1.52)	Moderate (D:/I:) p
Citalopram	Reboxetine	1.16 (0.29, 4.63)	Low 2,p	1.60 (1.08, 2.39)	High	1.60 (1.18, 2.15)	High (I:)
Citalopram	Sertraline	1.49 (1.02, 2.18)	High	0.84 (0.64, 1.09)	Moderate p	0.99 (0.78, 1.23)	Low (D:/I:) p,6
Citalopram	Venlafaxine	NA	NA	1.19 (0.95, 1.50)	Low 5,p	1.19 (0.95, 1.50)	Low (I:5) p
Duloxetine	Escitalopram	0.52 (0.26, 1.01)	Low 2,p	0.91 (0.62, 1.33)	Very low (3)5,p	0.70 (0.54, 0.92)	Low (D:2) 6
Duloxetine	Fluvoxamine	NA	NA	1.02 (0.71, 1.48)	Very low (3)5,p	1.02 (0.71, 1.48)	Very low (I:(3)5) p
Duloxetine	Milnacipran	NA	NA	0.86 (0.58, 1.29)	Low (3)p	0.86 (0.58, 1.29)	Low (I:(3)) p
Duloxetine	Mirtazapine	NA	NA	0.87 (0.62, 1.21)	Low (3/4)p	0.87 (0.62, 1.21)	Low (I:(3/4)) p
Duloxetine	Paroxetine	1.10 (0.81, 1.50)	Low 3,p	0.74 (0.49, 1.08)	Very low (2)5,p	0.92 (0.71, 1.20)	Low (D:3) p
Duloxetine	Reboxetine	NA	NA	1.20 (0.81, 1.75)	Low (3)p	1.20 (0.81, 1.75)	Low (I:(3)) p
Duloxetine	Sertraline	NA	NA	0.74 (0.54, 0.99)	Low (2/3)5	0.74 (0.54, 0.99)	Low (I:(2/3)5)
Duloxetine	Venlafaxine	NA	NA	0.89 (0.66, 1.20)	Very low (3)5,p	0.89 (0.66, 1.20)	Very low (I:(3)5) p
Escitalopram	Fluvoxamine	NA	NA	1.46 (1.07, 2.00)	High	1.46 (1.07, 2.00)	High (I:)
Escitalopram	Milnacipran	NA	NA	1.23 (0.87, 1.73)	Low 5,p	1.23 (0.87, 1.73)	Low (I:5) p
Escitalopram	Mirtazapine	NA	NA	1.23 (0.94, 1.60)	Very low (4)5,p	1.23 (0.94, 1.60)	Very low (I:(4)5) p
Escitalopram	Paroxetine	1.33 (0.85, 2.07)	Moderate p	1.31 (1.06, 1.62)	Moderate 5	1.31 (1.08, 1.59)	High (D:)
Escitalopram	Reboxetine	NA	NA	1.71 (1.23, 2.33)	Low (2)5	1.71 (1.23, 2.33)	Low (I:(2)5)
Escitalopram	Sertraline	0.81 (0.51, 1.29)	Low 3,p	1.11 (0.88, 1.42)	Moderate p	1.05 (0.84, 1.30)	Moderate (I:) p
Escitalopram	Venlafaxine	1.12 (0.72, 1.73)	Moderate p	1.30 (1.04, 1.64)	High	1.27 (1.03, 1.56)	High (D:/I:)
Fluvoxamine	Milnacipran	0.82 (0.36, 1.86)	Moderate p	0.85 (0.56, 1.30)	Very low (3)5,p	0.84 (0.58, 1.24)	Moderate (D:) p

(*Continued*)

Table 2. Continued

Treatment 1	Treatment 2	Direct Est.	Certainty	Indirect Est.	Certainty	Network Est.	Certainty
Fluvoxamine	Mirtazapine	1.20 (0.75, 1.93)	Low 1,p	0.72 (0.50, 1.03)	Low (4)p	0.84 (0.62, 1.14)	Low (D:1/I:(4)) p
Fluvoxamine	Paroxetine	0.93 (0.54, 1.61)	Moderate p	0.89 (0.64, 1.22)	Low (1/4)p	0.90 (0.68, 1.19)	Moderate (D:) p
Fluvoxamine	Reboxetine	NA	NA	1.17 (0.79, 1.71)	Moderate p	1.17 (0.79, 1.71)	Moderate (I:) p
Fluvoxamine	Sertraline	0.68 (0.09, 5.15)	Low 2,p	0.78 (0.56, 1.06)	Moderate p	0.72 (0.53, 0.97)	High (I:)
Fluvoxamine	Venlafaxine	0.49 (0.21, 1.18)	Very low 1,3,p	0.92 (0.68, 1.26)	Moderate p	0.87 (0.65, 1.16)	Moderate (I:) p
Milnacipran	Mirtazapine	NA	NA	1.00 (0.69, 1.45)	Low (4)p	1.00 (0.69, 1.45)	Low (I:(4)) p
Milnacipran	Paroxetine	1.14 (0.65, 1.99)	Moderate p	1.05 (0.72, 1.53)	Very low (3)5,p	1.07 (0.78, 1.47)	Moderate (D:) p
Milnacipran	Reboxetine	NA	NA	1.39 (0.91, 2.08)	Low (3)p	1.39 (0.91, 2.08)	Low (I:(3)) p
Milnacipran	Sertraline	0.59 (0.20, 1.74)	Low 1,p	0.89 (0.61, 1.27)	Very low (3)5,p	0.86 (0.60, 1.19)	Low (D:1) p
Milnacipran	Venlafaxine	NA	NA	1.03 (0.74, 1.45)	Very low (3)5,p	1.03 (0.74, 1.45)	Very low (I:(3)5) p
Mirtazapine	Paroxetine	1.19 (0.86, 1.65)	Low 4,p	1.00 (0.75, 1.34)	Low (4)p	1.07 (0.85, 1.34)	Low (D:4/I:(4)) p
Mirtazapine	Reboxetine	NA	NA	1.38 (0.97, 1.98)	Moderate p	1.38 (0.97, 1.98)	Moderate (I:) p
Mirtazapine	Sertraline	0.76 (0.45, 1.28)	Moderate p	0.87 (0.66, 1.16)	Moderate p	0.86 (0.66, 1.10)	Moderate (D:/I:) p
Mirtazapine	Venlafaxine	1.51 (0.99, 2.29)	Low 4,p	0.89 (0.67, 1.18)	Low (4)p	1.03 (0.81, 1.32)	Very low (D:4/I:(4)) p,6
Paroxetine	Reboxetine	NA	NA	1.30 (0.95, 1.75)	Moderate p	1.30 (0.95, 1.75)	Moderate (I:) p
Paroxetine	Sertraline	0.65 (0.27, 1.59)	Low 2,p	0.80 (0.64, 1.00)	Moderate p	0.80 (0.65, 0.97)	High (I:)
Paroxetine	Venlafaxine	0.84 (0.53, 1.33)	Moderate p	0.98 (0.81, 1.20)	High	0.97 (0.80, 1.16)	Moderate (D:/I:) p
Reboxetine	Sertraline	0.57 (0.12, 2.71)	Very low 1,3,p	0.62 (0.44, 0.85)	High	0.62 (0.45, 0.84)	High (I:)
Reboxetine	Venlafaxine	1.16 (0.38, 3.58)	Low 1,p	0.72 (0.52, 1.00)	Low 5,p	0.74 (0.55, 1.02)	Low (D:1/I:5) p
Sertraline	Venlafaxine	1.79 (0.73, 4.41)	Very low 1,2,p	1.12 (0.90, 1.41)	Moderate p	1.21 (0.99, 1.49)	Moderate (I:) p

In this table, the direct and indirect certainty ratings have the precision domain reincluded for fair comparison with network certainty ratings. In contrast, Appendix Table 2A shows certainty ratings for direct and indirect estimates that were used to inform (precision excluded) the certainty of the network estimate

NA = not applicable; there were no studies directly comparing the two treatments.

1 = Rated down due to issues with risk of bias.

2 = Rated down due to issues with consistency.

3 = Rated down due to issues with directness.

4 = Rated down due to issues with publication bias.

5 = Rated down due to issues with transitivity.

p = Rated down due to issues with precision; note that any rating down due to precision was done as a final step for direct, indirect, and network estimates.

6 = Rated down due to issues with coherence; 6* = evidence of incoherence but we did not rate down because the network estimate certainty was informed by an indirect estimate which had already been rated down for intransitivity.

(#) = Contributing evidence was rated down due to issues with #, which affects the rating for the current estimate.

D: = Network estimate rating initiated by rating of direct estimate.

I: = Network estimate rating initiated by rating of indirect estimate.

#/# = Used to show down-ratings when two contributing evidence sources had the same certainty rating and therefore both governed the initial rating for the current estimate.

Example in Indirect Est. column: (3,4/1,4) means both pairwise comparisons in the most dominant first order loop had the same certainty rating; the first comparison was rated down due to issues with directness and publication bias, and the second comparison was rated down for issues with risk of bias and publication bias.

Example in Network Est. column: (D:1,2/I:(1)5) means both direct estimate and indirect estimate had the same certainty rating; the direct estimate was rated down for issues with risk of bias and consistency, and the indirect estimate was rated down for issues with risk of bias and transitivity.

venlafaxine. One treatment was significantly better tolerated than the other in 13 (19.7%) of the 66 pairwise comparisons.

For acceptability, using our criteria, we found that 10 (15.2%) of the 66 pairwise comparisons had one treatment that was better tolerated than another. Escitalopram was better tolerated than duloxetine, fluvoxamine, paroxetine, and venlafaxine; and sertraline was better tolerated than fluvoxamine, paroxetine, and reboxetine. Owing to considering the certainty of the evidence, our conclusions about the relative tolerability of one treatment against another differed in 3 (23.1%) of 13 comparisons declared significant by the original authors.

3.4. Certainty of the evidence for the treatments ranked first

Mirtazapine, escitalopram, venlafaxine, and sertraline were the antidepressants with the highest estimated probability of being the most effective treatments. Their surface under the cumulative ranking curve (SUCRA) [44]

values were 91.9%, 87.1%, 81.9%, and 77.3%, respectively. The certainty of the evidence was low or very low in 15 (39.5%) of the 38 pairwise comparisons involving these treatments.

Escitalopram, sertraline, bupropion, and citalopram were the antidepressants with the highest estimated probability of being the best-tolerated treatments. Their surface under the cumulative ranking curve (SUCRA) values were 89.7%, 81.4%, 77.9%, and 77.6%, respectively. The certainty of the evidence was low or very low in 16 (42.1%) of the 38 pairwise comparisons involving these treatments.

Finally, considering the relative estimates and rank probabilities of efficacy and acceptability, as well as cost, Cipriani et al. [17] concluded that sertraline may be the preferred choice in patients with moderate to severe depression. When considering these elements along with the certainty of the evidence, sertraline was convincingly more effective than 3 of the 11 other alternatives and convincingly better tolerated to 3 of the 11 other treatments.

4. Discussion

In this study, to address the impact of adding GRADE certainty assessments to an NMA, we reanalyzed data from a published NMA and assessed the certainty of evidence from each pairwise comparison estimate using the GRADE approach. We found that the certainty of the estimates varied across pairwise comparisons and outcomes and modified the interpretation of the results in important ways. Specifically, by requiring moderate or high certainty evidence from network estimates to confidently conclude superiority of one treatment over another, the frequency of such inferences decreased to a more conservative amount. Of the 25 comparisons in which the original authors inferred superiority in terms of efficacy, 18 met our criteria for such a claim. For the outcome of acceptability, of the 13 comparisons that the authors of the NMA described as better tolerant, 10 met our criteria.

Our study has several strengths. First strength is our use of the GRADE approach that provides a comprehensive framework for assessing the certainty in a body of evidence. With extensive published guidance for its application, it has become the standard in rating certainty of evidence in systematic reviews with traditional head to head comparisons [45]. We applied the approach that the GRADE working group has suggested to extend to NMA, including some improvements to the originally published approach recently adopted by the GRADE working group [16]. Second, we reanalyzed the data to obtain all the information necessary to assess the certainty of the evidence from the NMA (that is, direct, indirect, and network estimates which are usually not provided in the articles reporting the results from NMAs). Third, all final judgments of risk of bias and certainty of the evidence were made after pairs of independent reviewers performed these assessments in duplicate and achieved consensus after discussion. Finally, a content expert who participated in the original NMA publication (TAF) both ensured that our clinical judgments were appropriate and clarified issues of uncertainty regarding the methods and results of the original NMA.

Some limitations of our study are intrinsic to the process of assessing the certainty of the evidence. We standardized all our assessments by establishing rules for judging the risk of bias at the study and at the outcome level and rules to determine when to rate down the certainty of the evidence. Inevitably, there are two limitations to this process. First, there is some arbitrariness to the specific criteria. Second, the rules do not obviate the need for judgment in their application. One example relates to our criteria for publication bias: we rated down certainty when all studies within a comparison were published and presented evidence of superiority for the antidepressant in which most studies were funded by the company producing the drug. Had a new treatment truly been successful, the evidence would exhibit such trends, and thus our judgment could be deemed unfair. With such few studies per pairwise comparison, statistical methods for inspecting the presence of publication bias were inappropriate to use [40], leaving us with this simplified criteria and the accompanying limitations.

A key judgment in NMA is the presence or absence of intransitivity. The judgment depends on inferences regarding effect modification, and the threshold for this judgment remains controversial. The authors of the original NMA were stringent in excluding studies with clear effect modification (for instance, studies including patients with bipolar disorder), leaving dose, age, and anxiety as the only plausible—though far from clearly established effect modifiers [46–48]. A different threshold for the judgments (i.e., not considering these features as effect modifiers) would have led to less, or possibly no, rating down for intransitivity.

Readers should in no way construe our findings as an adverse criticism of the work of Cipriani et al. [17]: their NMA was published before the availability of guidance on assessment of the certainty in the evidence from NMA. Moreover, ours is not the only available approach to applying GRADE to address the certainty of evidence from NMA. Other authors have developed an alternative approach involving a ‘‘contribution matrix’’ to evaluate study limitations during the process [26,49]. The extent to which the GRADE working group approach yields similar results to this alternative remains uncertain.

We judged up to half of the NMA estimates to have low or very low certainty, which is what determined the differences between Cipriani et al’s [17] conclusions and ours. The original review authors considered the risk of bias when drawing their conclusions implicitly, by including in their NMA only randomized trials that they judged to be at low risk of bias. We rated down 14 direct comparisons in the efficacy analysis and 5 in the acceptability analysis

for risk of bias due to limitations in concealment, blinding, and loss to follow-up (the difference in proportion rated down is that those who discontinued therapy and were then lost had events in the acceptability analysis but were lost in the efficacy analysis). While the original reviewer authors considered all incoherence likely to be explained by chance, a sensible conclusion, we opted to be conservative and, for the 4 (efficacy) and 5 (acceptability) pairwise comparisons in which incoherence was apparent, considered rating down certainty in the network estimate as well as the option of using direct or indirect estimates instead. As well, in the original review, there was no discussion around inconsistency between the trials in each pairwise comparison, indirectness, publication bias, intransitivity, and, aside from statistical significance, imprecision around the estimates of effect.

In contrast to our retrospective application, the current GRADE approach to NMA is best used by authors at the time of their review, while they are extremely familiar with all the data to make the necessary judgments. We strongly encourage review authors to, for each paired comparison, rate and publish the certainty in direct, indirect, and network estimates. Doing so will illuminate strengths and gaps in the evidence base and guide future research efforts.

Although, in this example, using the certainty of the evidence called for more circumspection in claiming treatment superiority, the extent to which this would be the case in other NMAs remains uncertain. It seems inevitable that application of GRADE will lead to recognition of comparisons supported by low or very low quality evidence and thus, more cautious inferences regarding the relative merits of the candidate interventions. Whatever the impact of application of GRADE in individual instances, it is bound to add value to any NMA. For instance, situations in which most evidence is of moderate to high certainty, conclusions regarding the relative merit of interventions may change little by application of GRADE, but it will be extremely reassuring to know that these relative estimates are based on trustworthy evidence and can thus be applied with confidence in the clinical arena.

Supplementary data

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.jclinepi.2018.05.009.

References

[1] Guyatt G, Rennie D, Meade MO, Cook DJ. Users' guides to the medical literature: a manual for evidence-based clinical practice 2015.

[2] Borenstein M, Hedges LV, Higgins J, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1:97−111.

[3] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Hoboken, NJ: John Wiley & Sons Inc; 2009.

[4] Sutton AJ, Higgins JP. Recent developments in meta-analysis. Stat Med 2008;27:625−50. https://doi.org/10.1002/sim.2934.

[5] Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. Heal Technol Assess 2005;9. 1−134, iii−iv.

[6] Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008;26:753−67.

[7] Nikolakopoulou A, Chaimani A, Veroniki AA, Vasiliadis HS, Schmid CH, Salanti G. Characteristics of networks of interventions: a description of a database of 186 published networks. PLoS One 2014;9:e86754.

[8] Bafeta A, Trinquart L, Seror R, Ravaud P. Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. BMJ 2013;347:f3675.

[9] Mills EJ, Ioannidis JPA, Thorlund K, Schünemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis multiple treatment comparison meta-analysis. JAMA 2012;308:1246−53.

[10] Miura T, Noma H, Furukawa TA, Mitsuyasu H, Tanaka S, Stockton S, et al. Comparative efficacy and tolerability of pharmacological treatments in the maintenance treatment of bipolar disorder: a systematic review and network meta-analysis. Lancet Psychiatry 2014;1:351−9.

[11] Zarin W, Veroniki AA, Nincic V, Vafaei A, Reynen E, Motiwala SS, et al. Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. BMC Med 2017;15:3.

[12] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.

[13] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ 2008;336:995−8.

[14] Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;64:380−2.

[15] Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014;349:g5630.

[16] Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk R, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol 2017. [Submitted].

[17] Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 2009;373:746−58.

[18] Gartlehner G, Gaynes BN, Hansen RA, Lohr KN. Ranking antidepressants. Lancet 2009;373:1761. author reply 1761-2.

[19] Ioannidis JP. Ranking antidepressants. Lancet 2009;373:1752−9.

[20] Jefferson T. Ranking antidepressants. Lancet 2009;373:1759. author reply 1761-2.

[21] Schwan S, Hallberg P. Ranking antidepressants. Lancet 2009;373:1761. author reply 1761-2.

[22] Seyringer ME, Kasper S. Ranking antidepressants. Lancet 2009;373:1760−2.

[23] Turner E, Moreno SG, Sutton AJ. Ranking antidepressants. Lancet 2009;373:1760. author reply 1761-2.

[24] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[25] The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011] 2011. Available from www.cochrane-handbook.org.

[26] Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. PLoS One 2014;9.

[27] Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux L, Van Noord M, et al. Comparative benefits and harms of second-

generation antidepressants for treating major depressive disorder: an updated meta-analysis. Ann Intern Med 2011;155:772−85.
[28] Gartlehner G, Hansen RA, Thieda P, DeVeaugh-Geiss AM, Gaynes BN, Krebs EE, et al. AHRQ Comparative Effectiveness Reviews. Comp. Eff. Second. Antidepressants Pharmacol. Treat. Adult Depress. Rockville (MD): Agency for Healthcare Research and Quality (US); 2007.
[29] Hansen RA, Moore CG, Dusetzina SB, Leinwand BI, Gartlehner G, Gaynes BN. Controlling for drug dose in systematic review and meta-analysis: a case study of the effect of antidepressant dose. Med Decis Mak 2009;29:91−103.
[30] Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Mak 2013;33:607−17.
[31] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010;29:932−44.
[32] van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Res Synth Methods 2016;7:80−93.
[33] van Vallkenhoef G, Lu G, de Brock B, Hillege H, Ades AE, Welton NJ. Automating network meta-analysis. Res Synth Methods 2012;3:285−99.
[34] R: A language and environment for statistical computing 2016.
[35] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.
[36] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.
[37] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence−study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.
[38] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence−inconsistency. J Clin Epidemiol 2011;64:1294−302.
[39] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence−indirectness. J Clin Epidemiol 2011;64:1303−10.
[40] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence−publication bias. J Clin Epidemiol 2011;64:1277−82.
[41] GH G, JW B. Modification of Cochrane Tool to assess risk of bias in randomized trials 2016.
[42] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683−91.
[43] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010;29:932−44.
[44] Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol 2011;64:163−71.
[45] Higgins JP, Lasserson T, Chandler J, Tovey D, R C. Methodologcal expectations of cochrane interventions reviews 2016.
[46] Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Ebert DD, et al. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiol Psychiatr Sci 2017;26:22−36.
[47] Bollini P, Pampallona S, Tibaldi G, Kupelnick B, Munizza C. Effectiveness of antidepressants. Meta-analysis of dose-effect relationships in randomised clinical trials. Br J Psychiatry 1999;174:297−303.
[48] Jakubovski E, Varigonda AL, Freemantle N, Taylor MJ, Bloch MH. Systematic review and meta-analysis: dose-response relationship of selective serotonin reuptake inhibitors in major depressive disorder. Am J Psychiatry 2016;173:174−83.
[49] Furukawa TA, Miura T, Chaimani A, Leucht S, Cipriani A, Noma H, et al. Using the contribution matrix to evaluate complex study limitations in a network meta-analysis: a case study of bipolar maintenance pharmacotherapy review. BMC Res Notes 2016;9:218.