Université Paris Cité

École doctorale 393 Pierre Louis de Santé Publique : Épidémiologie et Sciences de l'Information Biomédicale

Centre de Recherche en Épidémiologie et StatistiqueS (CRESS) (UMR 1153)
Équipe METHODS

Dealing with sparse data in network meta-analysis

Par Theodoros Evrenoglou

Thèse de doctorat de Biostatistiques

Dirigée par Anna CHAIMANI

Présentée et soutenue publiquement le 30 Novembre 2022

Devant un jury composé de :

Tim FRIEDE, Professor, Université de Gottingue, rapporteur

Stefan MICHIELS, DR-HDR, Université Paris Saclay, rapporteur

Nicky WELTON, Professor, Université de Bristol, examinatrice

Georgia SALANTI, Associate professor, Université de Berne, examinatrice

Raphaël PORCHER, Professeur des Universités, Université Paris Cité, examinateur

Anna CHAIMANI, CR-HDR, Université Paris Cité, directrice de thèse

Abstract in English

Title: Dealing with sparse data in network meta-analysis

In meta-analysis, the issue of sparse data can be defined as either the case where the studied outcome is rare across a set of available studies or as the case where there are only a few studies available. In both of those cases of sparsity the standard inverse-variance (IV) approach which is used for either pairwise or network meta-analysis can be problematic. This arises from the fact that in such cases the standard normal assumptions made by the conventional IV model are invalid and thus the summary effect estimates can be biased. This thesis aims to deal with sparse data through three different projects. The aim of the first project was to deal with the problem of rare events in NMA of binary data. An approach of penalizing the likelihood function has been previously proposed for bias reduction in the analysis of individual studies with rare events. To improve the accuracy and the precision of the NMA estimates in the presence of rare events, the first project aimed to extend the penalized likelihood approach to the context of NMA. The latter took place at first for the common-effect NMA model that uses logistic regression. The common-effect model was extended to a random-effects model by using a two-stage approach that incorporates the heterogeneity parameter through a multiplicative term. The performance of the PL-NMA model was evaluated through a simulation study of in total 33 scenarios where the method was also compared with 10 other NMA approaches in terms of various measures such as bias and coverage probability. The simulations suggested that the PL-NMA approach performs consistently well across all tested scenarios and most often results in smaller bias than the other NMA methods. The second project of this thesis aimed to facilitate the estimation of summary intervention effects for rare outcomes in the context of Covid-19 pandemic. The COVID-NMA initiative is a living

evidence synthesis platform that provides public access to the most up-to-date information with respect to the effects of the different Covid-19 interventions. Rare events are quite frequent in the COVID-NMA database. For example, across 432 RCTs with hospitalized patients in the database the median risk for the outcome of serious adverse events is only 8%. To facilitate the analysis of COVID-NMA data the metaCOVID application was built. This is a freely available R-Shiny application that allows all the end-users of the COVID-NMA platform to perform complex types of analysis through a user-friendly environment. metaCOVID deals with rare events by allowing the replacement of the IV model with other more suitable models for sparse data such as the Mantel-Haenszel or the PL-NMA model. The third project addressed the issue of sparse networks in NMA. This is the case of networks that contain a limited number of direct comparisons and very few studies to inform them. Results from such networks are accompanied with substantial uncertainty not only in NMA estimates but also in the plausibility of the underlying NMA assumptions. A Bayesian framework was proposed to allow sharing information between two networks that pertain to different subgroups of the population. Specifically, the results from a subgroup with a lot of direct evidence forming a 'dense' network were used to create informative priors for the relative effects of the target subgroup forming a sparse network. This is a two-stage approach where at the first-stage the results of the dense network are extrapolated to the sparse network using a NMA model which uses a location parameter that shifts the distribution of the relative effects to make them applicable to the target population. At the second-stage, the predictions of these results are used as prior information for the sparse network. The location parameter is informed either through the data or using expert opinion. The two-stage approach was found to result in more precise and robust estimates of the final NMA estimates.

Keywords: rare endpoints, bias reduction, multiple treatment meta-analysis, dynamic evidence synthesis; medical decision making, sparse networks, informative priors.

Résumé court en français

Titre: Gestion des données eparses dans les méta-analyses en réseau

En méta-analyse, le problème des données éparses peut être défini comme le cas où l'événement étudié est rare dans un ensemble d'études disponibles ou comme le cas où il y a peu d'études disponibles. Dans ces deux contextes, l'approche classique de la variance inverse (IV), utilisée pour la méta-analyse par paires ou par réseau (NMA), peut être problématique. Cela est dû aux hypothèses de normalités classiques faites par le modèle IV conventionnel qui ne sont pas valides et par conséquent les estimations de l'effet moyen peuvent être biaisées. Cette thèse se compose de trois projets différents. L'objectif du premier projet était de traiter le problème des événements rares dans la NMA dans le cadre de données binaires. Une méthode de pénalisation de la fonction de vraisemblance a été proposée précédemment pour réduire le biais dans l'analyse des études individuelles avec des événements rares. Afin d'améliorer l'exactitude et la précision des estimations de la NMA en présence d'événements rares, le premier projet visait à étendre l'approche de la fonction de vraisemblance pénalisée au contexte de NMA. Cette dernière a eu lieu, premiérement, pour le modèle NMA à effet commun qui utilise la régression logistique. Le modèle à effet commun a été étendu à un modèle à effet aléatoire en utilisant une approche en deux étapes qui incorpore le paramètre d'hétérogénéité par un terme multiplicatif. Les performances du modèle PL-NMA ont été évaluées grâce à une simulation comprenant 33 scénarios. La simulation a montré que l'approche PL-NMA produit de bons résultats dans tous les scénarios testés et résulte le plus

souvent en un plus petit biais que les autres méthodes de NMA. Le second projet de cette thèse visait à faciliter l'estimation des effets moyens des interventions pour des évènements rares dans le contexte de la pandémie de Covid-19. L'initiative COVID-NMA est une plateforme dynamique de synthèse de données qui offre un accès public aux informations les plus récentes sur les effets des différentes interventions contre le Covid-19. Les événements rares sont assez fréquents dans la base de données COVID-NMA. Par exemple, sur les 432 essais cliniques randomisés avec des patients hospitalisés dans la base de données, le risque médian pour l'événement d'effets indésirables graves est seulement de 8%. L'application metaCOVID est une application R-Shiny disponible gratuitement et qui permet à n'importe quel utilisateur de la plateforme COVID-NMA d'effectuer différentes analyses complexes de façon intuitive. metaCOVID traite les événements rares en permettant le remplacement du modèle IV par des modèles plus adaptés aux données éparses. Enfin, le troisième projet s'est penché sur la question des réseaux éparses dans la NMA. C'est le cas des réseaux qui contiennent un nombre limité de comparaisons directes et très peu d'études pour apporter des informations à ces comparaisons. Les résultats de tels réseaux s'accompagnent d'une grande incertitude, non seulement dans les estimations, mais aussi dans la plausibilité des hypothèses sous-jacentes de la NMA. Un cadre bayésien a été proposé pour permettre le partage d'informations entre deux réseaux qui se rapportent à différents sous-groupes de la population. En particulier, les résultats d'un sous-groupe comportant beaucoup de comparaisons formant un réseau "dense" ont été utilisés pour créer des informations a priori sur les effets relatifs du sous-groupe cible formant un réseau épars. Il s'agit d'une approche en deux étapes où, premièrement, les résultats du réseau dense sont extrapolés au réseau épars à l'aide d'un modèle NMA utilisant un paramètre de localisation qui déplace la distribution des effets relatifs pour les rendre applicables à la population cible. Deuxièment, les prédictions de ces résultats sont

utilisées comme prior pour le réseau épars. Il a été constaté que l'approche en deux étapes permettait d'obtenir des estimations plus précises et plus robustes des estimations NMA.

Mots clefs : critères de jugement rares, réduction de biais, méta-analyse de traitements multiples, synthèse dynamique des données, prise de décision médicale, réseaux épars, prior informatifs

Résumé substantiel en français

Les données éparses sont un problème fréquent et complexe qui peut affecter la validité de l'analyse statistique dans de nombreux domaines de la médecine factuelle. Même si cette thèse porte sur le domaine des revues systématiques et des méta-analyses, il convient de mentionner que le problème des données éparses peut également apparaître au niveau des études individuelles. Pour les études individuelles, les données éparses peuvent se présenter de deux manières différentes, soit dans le cas où la taille de l'échantillon disponible est petite, ce qui implique un faible nombre de participants, soit dans le cas où l'événement étudié est rare. Le premier cas, celui de la petite taille de l'échantillon, peut être plus général car il se produit avec n'importe quel type de données, tandis que le problème des événements rares apparaît seulement dans le cas de résultats dichotomiques.

Les estimations de l'effet de l'intervention dans les études individuelles sont souvent calculées en utilisant la théorie du maximum de vraisemblance. Les estimations du maximum de vraisemblance ont des propriétés asymptotiques souhaitables, qui suggèrent que lorsque la taille de l'échantillon augmente, les estimations du maximum de vraisemblance deviennent non biaisées. Par conséquent, on s'attend à ce que les estimations obtenues à partir d'études individuelles soient biaisées et imprécises lorsque les données sont éparses. Il est intéressant de noter que la présence de données éparses est fréquente pour les études individuelles. Un examen des revues disponibles sur PubMed a montré que la taille moyenne de l'échantillon d'un essai contrôlé randomisé est faible et se limite à 80 participants seulement. En outre, les événements rares peuvent également constituer un problème fréquent, notamment pour l'analyse de la tolérance ou des critères de jugement secondaires. Une étude empirique a montré que dans un échantillon de 1613 méta-

analyses de résultats de tolérance, environ 50% d'entre elles contenaient des essais avec un taux d'événements inférieur à 5%.

Les problèmes liés au biais et à la précision des estimations de l'effet des interventions dans les études individuelles peuvent également être rencontrés dans le cadre d'une méta-analyse. Dans les méta-analyses, le problème des données éparses se pose de la même manière que dans les études individuelles, soit en raison de la petite taille des échantillons, soit en raison de la rareté des événements. Il convient de faire une distinction importante car, dans le cas de la méta-analyse, le terme "taille de l'échantillon" fait référence à la quantité totale d'études disponibles et non pas au nombre total de participants. Ainsi, la rareté due à une petite taille d'échantillon implique qu'il n'y a que quelques études disponibles. La question des événements rares est une extension directe des études individuelles et fait référence au cas où il y a un faible nombre d'événements observés dans les études disponibles.

Les deux cas de rareté mentionnés précédemment peuvent causer des problèmes importants au modèle conventionnel de méta-analyse par paires à variance inverse, car il repose sur des approximations de grands échantillons qui risquent fort d'être invalides lorsque les données sont rares. Le modèle à variance inverse est un modèle en deux étapes basé sur le contraste, dans lequel, à la première étape, les effets d'intervention des études individuelles sont extraits avec une mesure de leur incertitude (ex: la variance). La première hypothèse de distribution, à savoir l'hypothèse intra-étude, suggère que les effets d'intervention spécifiques à l'étude sont modélisés par une distribution normale, avec pour variances réelles, les variances extraites des études individuelles. La deuxième hypothèse de distribution, l'hypothèse inter-études, n'est faite que par un modèle à effets aléatoires et suppose que les moyennes réelles spécifiques à l'étude suivent une distribution normale avec une grande moyenne et une variance qui est interprétée comme l'hétérogénéité entre

les études disponibles. Le paramètre d'hétérogénéité vise à tenir compte des différences cliniques et méthodologiques qui existent inévitablement dans un ensemble d'études indépendantes.

La présence de données éparses dans une méta-analyse par paires peut affecter la validité des deux hypothèses de distribution faites par le modèle de variance inverse à effets aléatoires. Lorsque peu d'études sont disponibles, l'estimation du paramètre d'hétérogénéité dans l'hypothèse inter-études peut être difficile. Ce dernier point est soutenu par une simulation qui a comparé au total 16 méthodes différentes pour estimer le paramètre d'hétérogénéité. Cette simulation a montré que sur l'ensemble des 16 méthodes, aucune n'avait de très bonnes performances dans les scénarios où seules quelques études (c'est-à-dire <5) étaient disponibles. Les problèmes liés à l'estimation du paramètre d'hétérogénéité peuvent conduire à des intervalles de confiance problématiques pour l'estimation globale de l'effet de l'intervention, car sa variation totale dépend à la fois de la variation totale provenant des études et de l'hétérogénéité entre elles. Ainsi, une estimation incorrecte du paramètre d'hétérogénéité peut conduire à des intervalles de confiance inappropriés avec une probabilité de couverture qui s'éloigne du niveau nominal habituel de 95%.

Lorsque les données sont éparses sous la forme d'événements rares, les hypothèses intra et inter-études du modèle peuvent être problématiques. L'hypothèse intra-étude traite les variances spécifiques des études comme si elles étaient vraies, alors qu'en fait, ce ne sont que des estimations de la vraie variance des études. Cette hypothèse ne peut être considérée comme presque vraie que lorsque les études respectives sont relativement importantes et comportent un nombre suffisant d'événements observés. Par ailleurs, en ce qui concerne l'hypothèse relative à l'ensemble des études, les méthodes standard d'estimation du paramètre d'hétérogénéité sont souvent insuffisantes pour traiter les événements rares et sont généralement biaisées. Le problème des événements rares peut devenir plus évident en présence d'études qui rapportent des événements nuls dans un ou tous

les bras de traitement. Dans de tels cas, le calcul des effets de l'intervention spécifique à chaque étude, tels que les odds ratios ou les risques relatifs, peut être impossible car ces derniers incluent une division par zéro. Deux solutions simples à ce problème suggèrent soit l'utilisation d'une "correction de continuité" pour les études qui rapportent des événements nuls, soit le remplacement des odds ratios ou des risques relatifs par les différences de risque pour estimer les effets de l'intervention. Cependant, les deux choix mentionnés ci- dessus se sont avérés problématiques dans plusieurs simulations. Plus précisément, la solution de la correction de continuité s'est avérée ajouter un biais supplémentaire aux résultats, tandis que le choix de la différence de risque entraîne des estimations particulièrement imprécises. Enfin, le manuel de Cochrane suggère que les études sans événements observés soient exclues de l'analyse car elles ne contribuent pas à l'estimation des estimations globales. Cette solution a été critiquée car ces études peuvent apporter de l'information à travers la taille de leur échantillon.

La méta-analyse en réseau (NMA) est une extension de la méta-analyse par paires qui permet la comparaison simultanée d'un nombre quelconque de traitements différents en concurrence. Cette méthode est connue pour être plus performante que la méta-analyse par paires et donne généralement des estimations plus précises car elle tient compte des sources d'information directes et indirectes. Les avantages de la méta-analyse en réseau dépendent en grande partie de la validité de l'hypothèse de transitivité. Celle-ci garantit la validité des comparaisons indirectes dans le réseau en exigeant que la distribution des modificateurs d'effet soit la même dans toutes les comparaisons qui impliquent un comparateur commun. L'hypothèse de transitivité ne peut être vérifiée que qualitativement sur la base d'un avis clinique. Cependant, sa représentation statistique, à savoir l'hypothèse de consistance, peut être vérifiée pour tester les divergences potentielles entre les estimations directes et indirectes dans le réseau.

Les problèmes causés par les données éparses peuvent inévitablement survenir au niveau de la méta-analyse de réseau. Ceci est prévisible puisque le modèle conventionnel de méta-analyse en réseau est toujours le modèle de variance inverse. Pour la méta-analyse de réseau, le problème des données éparses apparaît soit sous la forme de réseaux épars, soit sous la forme d'événements rares. Les réseaux épars sont définis comme étant des réseaux qui contiennent un nombre limité de comparaisons directes et très peu d'études pour les renseigner. Les résultats de ces réseaux s'accompagnent d'une incertitude substantielle non seulement dans les estimations de la NMA mais aussi dans la plausibilité des hypothèses sous-jacentes de la NMA. Par exemple, vérifier la validité ou non de l'hypothèse de consistance dans les réseaux épars peut s'avérer difficile car les divergences potentielles entre les estimations directes et indirectes peuvent être masquées par la grande incertitude qui les accompagne. Les biais dus à des événements rares peuvent se produire dans les méta-analyses en réseau, car l'hypothèse du modèle de variance inverse intra et inter-études est à nouveau problématique. De plus, pour la méta-analyse en réseau, la question supplémentaire de la connectivité du réseau peut se poser en présence de nombreuses études rapportant des événements nuls dans tous les bras de traitement. Dans de tels cas, l'exclusion extensive des études avec zéros événements peut entraîner la déconnexion des réseaux initialement connectés.

Cette thèse vise à traiter le problème des données éparses à travers trois projets distincts. Le premier projet traite du problème des événements rares dans la méta-analyse de réseaux avec des données binaires. Une approche de pénalisation de la fonction de vraisemblance a été précédemment proposée par Firth pour réduire les biais dans l'analyse des études individuelles avec des événements rares. Pour améliorer l'exactitude et la précision des estimations de la NMA en présence d'événements rares, le premier projet visait à étendre l'approche de la fonction de

vraisemblance pénalisée au contexte de la NMA. Celui-ci a eu lieu dans un premier temps pour le modèle NMA à effet commun qui utilise la régression logistique. Il s'agit d'un modèle à une étape qui présente un avantage conceptuel par rapport au modèle à variance inverse à deux étapes, car il remplace l'approximation de la distribution normale dans les études par la distribution binomiale exacte. De plus, le modèle de vraisemblance pénalisé a l'avantage de permettre l'inclusion d'études avec zéros événements sans exiger de correction de continuité. Cela préserve la connectivité du réseau et conduit potentiellement à des résultats plus précis puisque toutes les études disponibles sont prises en compte pour le calcul des estimations globales. Le modèle à effets communs a été étendu à un modèle à effets aléatoires en utilisant une approche qui incorpore un paramètre d'hétérogénéité grâce à un terme multiplicatif.

La performance du nouveau modèle NMA à vraisemblance pénalisée a été évaluée par une simulation extensive de 33 scénarios au total. Dans cette simulation, l'approche proposée a été comparée à 10 autres modèles de NMA, dont le modèle conventionnel de variance inverse. Diverses mesures de performance telles que le biais moyen et la probabilité de couverture ont été utilisées pour comparer les différents modèles ajustés. Le modèle de vraisemblance pénalisé s'est avéré être plus performant que les autres modèles en termes de biais. Le modèle de variance inverse s'est avéré inadapté à l'analyse d'événements rares car les estimations résultantes étaient fortement biaisées. En termes de probabilité de couverture et d'erreur quadratique moyenne, tous les modèles ajustés ont montré des performances similaires, tandis que le modèle de vraisemblance pénalisé a fourni les résultats les plus précis, la longueur moyenne de l'intervalle de confiance obtenue par ce modèle étant plus petite. L'impact de l'inclusion ou de l'exclusion des études sans événement a été évalué dans la simulation. En particulier, le nouveau modèle de vraisemblance pénalisé a été ajusté une première fois en incluant et une seconde fois en excluant les études comportant zéro événement

dans tous les bras de traitement. Il est intéressant de noter que l'inclusion ou l'exclusion de ces études a eu un impact négligeable sur l'estimation ponctuelle des effets de l'intervention. Cependant, l'inclusion de ces études s'est avérée bénéfique en termes d'amélioration de la probabilité de couverture des intervalles de confiance de Wald. En ce qui concerne l'estimation du paramètre d'hétérogénéité, tous les modèles à effets aléatoires ajustés se sont avérés peu performants. Cela implique qu'en présence d'événements rares, des travaux supplémentaires sont requis pour identifier les méthodes appropriées d'estimation du paramètre d'hétérogénéité. Dans l'ensemble, la nouvelle approche s'est avérée être une solution fiable pour effectuer une méta-analyse de réseau en présence d'événements rares.

Le deuxième projet de cette thèse visait à faciliter l'estimation des effets globaux des interventions pour les événements rares dans le contexte de la pandémie de Covid-19. L'initiative COVID-NMA est une plateforme de synthèse de données dynamique qui fournit un accès public aux informations les plus récentes concernant les effets des différentes interventions contre le Covid-19. Les événements rares sont assez fréquents dans la base de données COVID-NMA. Par exemple, sur les 432 essais cliniques randomisés avec des patients hospitalisés dans la base de données, le risque médian pour le résultat des événements indésirables graves est seulement de 8%. L'application metaCOVID a été conçue afin de faciliter l'analyse des données COVID-NMA. Il s'agit d'une application R-Shiny disponible gratuitement qui permet à tous les utilisateurs de la plateforme COVID-NMA d'effectuer différentes analyses complexes dans un environnement simple d’utilisation. metaCOVID traite les événements rares en permettant le remplacement du modèle IV utilisé dans la plateforme COVID-NMA par d'autres modèles plus adaptés aux données éparses tels que le modèle de Mantel-Haenszel ou le modèle de vraisemblance pénalisé qui a été proposé dans le premier projet de la thèse.

Le troisième projet s'est penché sur la question des réseaux épars dans la méta-analyse de réseau. Un cadre bayésien a été proposé pour permettre le partage d'informations entre deux réseaux qui se rapportent à différents sous-groupes de la population. Plus précisément, les résultats d'un sous-groupe avec beaucoup de preuves directes formant un réseau "dense" ont été utilisés pour créer des priors informatifs pour les effets relatifs du sous-groupe cible formant un réseau épars. Il s'agit d'une approche en deux étapes où, à la première étape, les résultats du réseau dense sont extrapolés au réseau épars à l'aide d'un modèle NMA hiérarchique modifié utilisant un paramètre de localisation qui déplace la distribution des effets relatifs pour les rendre applicables à la population cible. Dans un deuxième temps, les prédictions de ces résultats sont utilisées comme prior pour le réseau épars. La nouvelle approche en deux étapes a été illustrée à travers un exemple de patients psychiatriques où le sous-groupe cible des enfants-adolescents forme un réseau épars et le sous-groupe des "patients généraux" un réseau dense. Le paramètre de localisation dans cet exemple illustratif a été renseigné soit à partir des données, soit en utilisant l'avis d'experts. Un questionnaire a été distribué à 22 cliniciens et experts dans le domaine de la schizophrénie. Les réponses obtenues par les experts ont facilité l'extrapolation qui a lieu à la première étape de la méthode. Dans l'ensemble, l'approche en deux étapes a permis d'obtenir des estimations beaucoup plus précises et robustes des effets relatifs, qui peuvent donner un aperçu de l'efficacité des antipsychotiques pour le sous-groupe cible et informer de manière adéquate la pratique clinique.

Acknowledgments

First and foremost, I would like to thank my supervisor Dr. Anna Chaimani for her endless support, continuous encouragement, important guidance and mentorship over all my PhD years.

I am grateful to Professor Isabelle Boutron for always being available to advise me and for giving me the great opportunity to increase my knowledge and expertise through my work in the COVID-NMA project.

I would like to thank all the members of the jury for their useful comments and remarks in my PhD thesis.

I owe special thanks to all the co-authors of my PhD thesis articles for their important contributions and inputs in my work. I would like to specifically thank Professor Ian R. White for his feedback in my first article and for accepting me to work close to him for one month at the MRC Clinical Trials Unit at University College London.

I would like to thank all the members of the METHODS lab for their constant support during my PhD studies. In particular, Professor Raphaël Porcher for his support that started even from the audition stage before I join the lab, Dr. Silvia Metelli for all the fruitful discussions and the feedback that she very generously gave me every time I needed, Florie Brion Bouvier and Alan Balendran for helping me with the French abstract of my thesis.

I thank Romina, Savvas, Anna, Eugenios and Maria for their gentle support and for offering me all these nice breaks which kept me focused on my work.

I thank my mother Marina and my brother Yiannis for always supporting me. Without them I would not have managed even half of my way.

List of abbreviations

RCT: Randomized Control Trial

OR: Odds Ratio

RR: Risk Ratio

GLM: Generalized Linear Model

MLE: Maximum Likelihood Estimate

IV: Inverse Variance

REML: Restricted Maximum Likelihood

DL: DerSimonian Laird

PM: Paul Mandel

RD: Risk Difference

MH: Mantel-Haenszel

HN: Half Normal

NMA: Network Meta Analysis

IF: Inconsistency Factor

SUCRA: Surface Under Cumulative Ranking

NCH: Non Central Hypergeometric

Covid-19: Coronavirus disease 19

PL-NMA: Penalized Likelihood Network Meta Analysis

1. Introduction

1.1. The problem of sparse data bias in individual studies

Sparse data are a common problem that can threaten the validity of the statistical analysis across various fields of evidence-based medicine[1,2]. Although the current thesis is dedicated to the level of evidence which refers to systematic reviews and meta-analyses, the problem of sparse data can be broader and it can also appear for individual studies. Sparse data can occur for individual studies either as the case of a small sample size or as the case where the studied outcome is rare[1]. The problem of a small sample size can exist with any type of data (e.g. continuous, binary etc.) as it is connected to the number of participants which are included in each study. The problem of rare events is connected to the frequency that the studied outcome is observed. The latter implies that rare events occur for outcome data which are modeled through categorical variables. In the current thesis rare events will always refer to the case where the studied outcome is binary[2].

Problems with spase data frequently appear when investigating sensitive subgroups of the population such as children and adolescents, elder patients or individuals with multimorbidity[3]. Each of the two forms of sparsity has previously been found to be quite frequent in the published literature. A review across PubMed journals has shown that the average sample size for a randomized controlled trial (RCT) is small and is limited to only 80 participants[4]. Additionally, rare events can be frequent especially for the analysis of safety or secondary outcomes. An

empirical study has found that in a sample of 1613 meta-analyses of safety outcomes around 50% of them were containing trials with event rate lower than 5%[5].

Effect measures such as odds ratios (OR) or risk ratios (RR) are usually calculated for individual studies while also adjusting for a set of different independent covariates. The latter is usually possible through the so-called generalized linear models (GLMs)[6,7]. Suppose that $Y_1, Y_2, \dots, Y_n$ are independent realizations from a random variable **Y**. A GLM can be parametrically written as,

$$g(\boldsymbol{\gamma}) = \mathbf{Z}\boldsymbol{\beta},$$

where $g(\boldsymbol{\gamma})$ is a link function that relates the mean value $\boldsymbol{\gamma}$ of the random variable $\boldsymbol{Y}$ with the linear predictor part which consists of the $n \times p$ model's design matrix $\mathbf{Z}$ and the parameter vector $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)^{\mathrm{T}}$. Different choices of a link function can lead to different types of regression models. For example, the logistic regression model for modeling binary outcome data can arise after the choice: $g(\boldsymbol{\gamma}) = log\left(\frac{\gamma}{1-\gamma}\right)$.

The estimation of the model's parameter vector $\boldsymbol{\beta}$ relies on the maximum likelihood theory. The maximum likelihood estimates (MLE) have a central role across all models in the class GLMs because they are simple to calculate as the maximum values of the respective likelihood function. Moreover, the method has attractive analytical properties such as that while the sample size increases the distribution of the parameter $\boldsymbol{\beta}$ converges to a normal distribution with mean value equal to the MLE $\widehat{\boldsymbol{\beta}}$. The latter implies that the MLEs are asymptotically unbiased. This unbiasedness of the MLE is valid only in the theoretical case where the sample size n goes to infinity[8]. In practical applications though the MLE is always expected to carry some amount of bias which can be expressed asymptotically as[8,9],

$$B(\boldsymbol{\beta}) = \frac{b_1(\boldsymbol{\beta})}{n} + \frac{b_2(\boldsymbol{\beta})}{n^2} + \cdots \quad (1.1)$$

where the functions $b_i(\cdot), i = 1,2,..n$ are bounded functions which depend on $\boldsymbol{\beta}$.

The expression in equation (1.1) suggests that the bias of the MLE is an additive term and is consistent with the already known property of the MLE to be asymptotically unbiased. The latter can be easily seen in (1.1) as when $n \to \infty$ then $B(\boldsymbol{\beta}) \to 0$. However, when the data are sparse then the most of the terms in equation (1.1) are not canceling out. Therefore, as the data become sparser the bias of the MLE increases. This kind of bias is usually termed in the literature as sparse data bias[1]. Beyond the bias of point estimates problems can also arise in terms of the precision of the MLE as those can be imprecise[10]. For example, in the case of logistic regression the variance of the MLE $\widehat{\boldsymbol{\beta}}$ is known to be proportional to the quantities $\frac{1}{\hat{p}_i(1-\hat{p}_i)}$, where p_i denotes the probability of event for the i^{th} individual. This variance expression suggests that as the outcome of interest becomes rarer the probability of event decreases thus yielding less precise MLE.

Finally, when MLEs refer to a logistic regression model then in the presence of sparse data the problem of separation can additionally occur[11,12]. As separation it is defined the case where one covariate or a linear combination of covariates perfectly predict the outcome of interest. In such cases, the likelihood function goes to infinity and yields to undefined MLEs[11]. A common example of separation refers to the case where 0 events are observed in both the intervention and the control group. In this case the data are completely separated thus leading to infinite MLEs[12] .

Overall, important statistical challenges can exist for individual studies when the data are sparse. In such cases the MLE estimates for commonly used effect measures such as OR and RR are known to have bias away from the null[1]. The magnitude of this bias can increase considering

the fact that the final estimates of those measures are calculated after the exponentiation of their initial logarithmic value. This can substantially worsen the situation and lead to estimates which suffer from further upward bias. Finally, for logistic regression models the bias of the MLE's can even become infinite in cases of separated data[11].

1.2. Basic concepts of pairwise meta-analysis

Systematic reviews and meta-analyses have been established as an integral part of comparative effectiveness research[13] and they are widely used by important healthcare organizations such as the World Health Organization or NICE. The goal of a meta-analysis is to synthesize the findings of individual studies and to summarize the effectiveness or safety of an experimental intervention compared to a control group[14]. This is the core advantage of meta-analysis over isolated individual studies which are often small and underpowered to provide reliable estimates about the relative effectiveness or safety of different interventions[15].

It is inevitable that across a set of independent individual studies there will always exist clinical and methodological differences[14]. Those differences might be reflected in the final summary intervention effect in the form of additional uncertainty beyond the one expected from the random-error. Such a type of uncertainty is known as heterogeneity[14,16–18] and it can be treated as a second source of variation. The different considerations about heterogeneity are important in meta-analysis as they lead into the two basic types of meta-analytical models: i) the common-effect model and ii) the random-effects model[17,19].

Consider a set of N individual studies which are included in a systematic review. Each study $i = 1,2, \dots N$ provides an observed relative effect y_i between two interventions 1 and 2. This observed relative effect is always accompanied with an observed variance s_i^2. The common-effect

model assumes that the study-specific true intervention effects θ_i are identical across the N studies, hence

$$\theta_i = \mu$$

where μ is the true summary relative effect. The only distributional assumption made by a common-effect model is that the observed relative effects y_i are following a normal distribution with mean μ and variance s_i^2, i.e. $y_i \sim \mathrm{N}(\mu, s_i^2)$. In this case the summary relative effect can be estimated as

$$\hat{\mu} = \frac{\sum_i w_i y_i}{\sum_i w_i}, i = 1,2, \dots, N \quad (1.2)$$

with w_i to denote the weights assigned for study i. Those are calculated as $w_i = \frac{1}{s_i^2}, i = 1,2, \dots n$. The weighting of the studies can also be done according to other criteria than the inverse of the study variances (IV)[14] . However, the IV approach is usually the most preferable as in this case the resulting estimate in equation (1.2) is a MLE and thus it carries the good asymptotic properties described in the previous section[20].

Even though the common-effect model allows for a simple way to calculate the summary relative effect, the model relies on the rather unrealistic assumption that the true intervention effects θ_i are identical across a set of N studies. The random-effects model[17] makes the more realistic assumption that for each study $i = 1,2, \dots, N$ there is a distinct true underlying intervention effect. The random-effects θ_i are assumed to share a common distribution which conventionally is the normal distribution[20]. The distributional assumption that models the study specific data will be termed as within-studies assumption and the distributional assumption that models the random-

effects θ_i will be termed as across-studies assumption. A random-effects model can be written by using a hierarchical model expression as,

$$y_i \sim N(\theta_i, s_i^2) \tag{1.3}$$

$$\theta_i \sim N(\mu, \tau^2)$$

where the variance parameter τ^2 represents the variation in the true study-specific intervention effects and it can be interpreted as the heterogeneity parameter. The summary relative effect of a random-effects meta-analysis model is estimated as,

$$\hat{\mu} = \frac{\sum_i w_i^* y_i}{\sum_i w_i^*} \tag{1.4}$$

where the weights w_i^* are now calculated as $w_i^* = \frac{1}{s_i^2 + \hat{\tau}^2}$. The only difference between equations (1.2) and (1.4) lies on the incorporation of the heterogeneity parameter in the study weights through the random-effects model. This means that by setting $\tau^2 = 0$ the random-effects model corresponds with the common-effect model. Finally, it should be noted that the choice of the normal likelihood for the within-studies assumption is only a conventional choice of a distribution. Alternatively, other types of distributions can be made such as the binomial or the Poisson distribution for modeling binary or count data, respectively. Such choices of likelihoods lead to the one-stage meta-analytical models which perform a meta-analysis by using GLMs[21].

Multiple methods have been proposed in the frequentist setting for the estimation of the parameter τ^2. Among the most common ones are the restricted maximum likelihood method (REML)[22], the Der Simonian and Laird (DL)[23] moment based estimate and the Paul-Mantel[24] (PM) estimate. The REML method is the most commonly used method as it has been found to outperform many other methods in terms of bias[25]. Alternatively, both common and random-

effect(s) meta-analytical models can be fitted in a Bayesian framework[26] where in this case prior distributions need to be assigned for the parameters of interest μ and τ^2.

The incorporation of the heterogeneity parameter through a random-effects model is only the first step in meta-analysis. In the presence of a substantial amount of heterogeneity, the final summary relative effect can be misleading and it may not properly represent some of the included individual studies[14]. Additional amounts of heterogeneity can be explained through the so-called meta-regression models[18,27,28]. Conceptually, a meta-regression model resembles a weighted linear regression model where the outcome variable refers to the observed study-specific effect estimates and the covariates are the study-characteristics which are judged to be influential on the intervention effect[14]. Parametrically, a meta-regression model can be written as,

$$y_i = \mu + \sum_{j=1}^{p} \beta_j z_{ij} + \theta_i + \varepsilon_i$$

with $\theta_i \sim \mathrm{N}(0, \tau^2)$ and $y_i \sim \mathrm{N}(0, s_i^2)$. The parameters $\beta_j, j = 1,2, \dots, p$ are the regression coefficients which are interpreted as: "how much a unit increase in the covariates $z_{ij}, i = 1,2, \dots, N, j = 1,2 \dots, p$ will alter the intervention effect". In the case that the covariates z_{ij} are properly chosen; it is then expected that the estimate of the heterogeneity parameter obtained from a meta-regression model will be smaller than the corresponding one obtained from a meta-analysis model without covariates.

1.3. The problem of sparse data in pairwise meta-analysis

Meta-analysis models can face important limitations when they are applied for sparse data. In meta-analysis the concept of sparse data appears as either the case of small sample sizes or as the case when the studied outcome is rare across the available studies[29]. An important distinction needs to be made as the term "sample size" here refers to the number of the available studies and is not directly related to the number of participants. The issue of performing a meta-analysis to a set of studies that contain a small number of participants can be another form of sparsity that may arise[30]. However, the latter goes beyond the scope of this thesis.

1.3.1. Conducting pairwise meta-analysis with few available studies

When performing a meta-analysis with few available studies statistical problems[31,32] can arise for the estimation of both μ and τ^2. As mentioned in the previous section the estimation of the parameter μ relies on the maximum likelihood theory[20] which is known to lead to biased estimates in cases of small sample sizes. Important issues can also arise in the estimation of the heterogeneity parameter τ^2[31,33]. The REML method is known to provide biased estimates when few studies are available (e.g. <5 studies)[34,35]. Moreover, the calculation of the REML estimate relies on an iterative process which can have convergence problems in cases with only few studies[34,35]. The problematic estimation of the heterogeneity parameter can affect the estimation of the confidence interval of the estimated summary effect $\hat{\mu}$[32]. This is usually calculated as

$$\hat{\mu} \pm Z * SE(\hat{\mu}),$$

where Z is a value arising from the Z-statistic and $SE(\hat{\mu})$ is the standard error of $\hat{\mu}$ calculated as $\frac{1}{\sqrt{w_i^*}}$. A biased estimate for τ^2 can lead to an improper study-specific weights w_i^* and as a consequence to problematic confidence intervals for $\hat{\mu}$ with coverage probability away from the usual nominal level of 95%.

Different alternatives to the REML method have been proposed in the literature as the problem of the poor estimation of τ^2 is a well-known issue. The DL[23] moment based method is an analytical approach for estimating τ^2 which is implemented in many software routines[36,37]. This approach has the advantage of always resulting in estimates for τ^2 as it does not rely on any iterative process. Simulation studies have shown that the DL performs considerable well only in cases when true value of parameter τ^2 is small or zero and the sample size is large[38–40]. However, the DL method can result in significantly biased estimates in cases where τ^2 is large. Another alternative is the iterative PM[24] method. This method has been found in many simulations to outperform DL in terms of bias[34,35,38,39]. The PM estimate is approximately unbiased when the sample size is large. Although, additional results have also shown that this estimate can suffer from upward bias in cases of small true values of τ^2 and small sample sizes. Overall, in a summary review Veroniki et al.[34] compared 16 different methods for estimating the heterogeneity parameter τ^2. The authors found that when few studies are available (e.g. <5) all the 16 methods were biased and thus they recommended that an extensive sensitivity analysis should always carried out by using multiple choices of estimation methods.

1.3.2.Conducting pairwise meta-analysis in presence of rare events

Individual studies are often underpowered to detect intervention effects in cases of rare events[41]. In such cases a meta-analysis of many studies can be the only reliable way to detect

whether an intervention impacts the occurrence of a rare outcome. However, the normal assumptions made by the IV model are only large sample approximations which are very likely to be invalid in cases of rare events[41–43].

The within-studies assumption in equation (1.3) suggests that the intervention effects y_i observed in study i can be modeled by a normal distribution with true means θ_i and 'true' variances the variances s_i^2 which are extracted from study i. The latter is a general source of criticism for the IV model[44] as in this way the within-study variances s_i^2 are treated as fixed and known whether in fact those are only estimates which are accompanied with an amount of uncertainty[45]. The failure of the IV model to account for the uncertainty of the within-study variances s_i^2 is expected to lead to the underestimation of the uncertainty of the overall intervention effect estimate $\hat{\mu}$ as a source of variation arising from the within-study variances is considered as negligible.

Important issues can also arise in terms of the across-studies assumption in equation (1.3) as again the estimation of the heterogeneity parameter can be biased. Specifically, the most of the proposed frequentist methods for estimating τ^2 usually underestimate this parameter as 0[41,46]. This renders the random-effects IV model to its common-effect counterpart. As a consequence, the source of variation coming from the heterogeneity across the studies is not taken into account thus leading to spuriously narrow confidence intervals for the summary estimate $\hat{\mu}$. Finally, it should be noted that even though the estimation of the heterogeneity parameter is a core step in every meta-analysis the Cochrane handbook suggests that in presence of rare events this should be treated as of secondary interest and the emphasis should always be given in the proper estimation of the summary intervention effect[14].

Problems with rare events can become more evident in the presence of studies that report zero events in one or all intervention arms[42,43]. In such cases the calculation of study-specific intervention effects such as the RR or the OR is impossible since this involves the division with zero. An easy way to overcome this problem can be the use of a 'continuity correction' for the studies which report zero events[42]. Let r_{ik} and n_{ik} denote the number of observed events and the sample size in arm $k = 1,2$ of study $i = 1,2 \dots, N$, respectively. The solution of the continuity correction suggests that a fixed value (e.g. 0.5) should be added to all the cells r_{ik} and $n_{ik} - r_{ik}$ of the study i which reports zero events in one or all intervention arms. This solution bypasses the problems caused by the zero events and allows to perform a meta-analysis using the IV model. Simulation studies have shown that applying a continuity correction for the IV model can lead to excess amounts of bias in the estimated intervention effects[43] and thus this solution is not recommended.

A second solution to the problems caused by zero event studies can be the replacement of the OR and the RR with the risk difference (RD). In this case the IV model can be applied without any of the computational problems. However, it has been found that when the events are rare the RD can lead into extremely wide confidence intervals with very low power[43]. Due to those poor statistical properties the choice of RD is unsuitable for tackling meta-analysis of rare events[43]. Finally, it is worth mentioning that Cochrane handbook suggests that studies with zero events in all of their intervention arms, namely double-zero studies, should be omitted from the meta-analysis as those studies do not contribute any information in the estimation of the summary intervention effect[14]. In that case though some researchers have argued that from an ethical point of view the patients who are included in double-zero studies deserve to be included in the meta-

analysis[47] while other researchers have also raised the issue that such studies do carry information through their sample sizes[48].

Rare events are well studied in the statistical literature of meta-analysis and different alternatives to the IV have been proposed to deal with this problem. The emphasis across the proposed methodologies is usually given on proper ways to avoid the within-studies normal assumption[45]. A common way to avoid this assumption is to replace the parametric IV model with other non-parametric common-effect approaches such as the Mantel-Haenszel (MH)[49] or the Peto methods[50]. The MH method is usually used for the estimation of a summary OR but it can alternatively be used for the estimation of either a summary RR or RD. The summary OR obtained by the MH can be calculated as,

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{N} \frac{r_{i2}(n_{i1} - r_{i1})}{n_{i+}}}{\sum_{i=1}^{N} \frac{r_{i1}(n_{i2} - r_{i2})}{n_{i+}}}$$

where n_{i+} denotes the total sample size in study i. The MH method can include single zero studies without any continuity correction unless the case where either r_{i1} or r_{i2} are equal to 0, $\forall i$. The Peto method can be used only for the estimation of a common OR. The summary log odds ratio (logOR) according to the Peto method can be calculated as,

$$\log\left(\widehat{OR}_{Peto}\right) = \frac{O - E}{V}$$

where $O = \sum_{i=1}^{N} r_{i2}$, $E = \sum_{i=1}^{N} \frac{n_{i2} r_{i+}}{n_{i+}}$ and $V = \frac{n_{i1} n_{i2} r_{i1} (n_{i1} - r_{i1})}{n_{i+}^2 (n_i - 1)}$ with r_{i+} being the total number of events in study i. This method is also a more flexible approach compared to the IV in terms of including studies with zeros. Peto's method can include single zero studies without any continuity correction while the method can have computational issues only in cases where all the available

studies are report zero events in both of their intervention arms. Simulation studies have shown that the use of Peto's method can provide nearly unbiased estimates only when the risks for events are very low and sample sizes are balanced across the available studies while in all other cases the method has been found to be biased[42,43].

Beyond the non-parametric approaches other parametric methods can be used for rare events instead of the IV model. Those methods aim to replace the problematic within-studies normal assumption with other more suitable choices of likelihoods. Logistic regression models can be used in meta-analysis as one-stage approaches[51]. Those models use the exact binomial distribution of the arm level data to model the observed events in study i as $r_{ik} \sim Bin(n_{ik}, p_{ik})$ with p_{ik} denote the probability of event in arm $k = 1,2$ of study $i = 1,2, \dots N$. A random-effects meta-analysis model that uses logistic regression can be parametrized as,

$$logit(p_{ik}) = \alpha_i + \delta_{i,1k}, i = 1,2, \dots N, k = 1,2 \quad (1.5)$$

$$\delta_{i,1k} \sim N(d_{1k}, \tau^2)$$

where α_i denotes the fixed study-specific intercepts which are interpreted as the log odds of event in group 1 of study i. The random-effects $\delta_{i,1k}$ denote the true study-specific logORs which are modeled by a normal distribution with grand mean the logOR d_{1k} and variance the heterogeneity parameter τ^2.

The estimation of the heterogeneity parameter τ^2 can be challenging in presence of rare events for the mixed model in equation (1.5)[41]. The latter is a general caveat of the model as the failure in the estimation typically results in $\hat{\tau}^2 = 0$ thus rendering the model (1.5) to its common-effect counterpart[46]. Single zero studies are taken into account by the logistic regression model in (1.5) while double zero studies are excluded from the analysis as in that case the data are perfectly

separated thus leading to infinite MLEs[11,12]. The inclusion of double-zero studies can be possible when the model in (1.5) is slightly modified to allow for random study intercepts (i.e. $\alpha_i \sim N(a, \sigma^2)$). This extension allows sharing information between the different studies thus allowing for the inclusion of double zero studies in the analysis[46,52].

The model in (1.5) can also be fitted in the Bayesian framework[53]. The latter requires the choice of a prior distribution for the parameters of interest a_i, d_{1k} and τ^2. A standard choice of prior for a_i can be the non-informative $N(0,10^2)$ distribution while a non-informative $N(0,100^2)$ is often used for the parameter d_{1k}. Although, for rare events the choice of the weakly informative priors $N(0, 2.82^2)$ and $HN(0.5)$ for d_{1k} and τ has been found to be beneficial for improving the performance of the model in terms of bias in comparison to the respective model that uses a non informative prior for d_{1k}[53].

Additional choices of models can be either the conditional logistic regression[54], the beta-binomial model[48], the Poisson-Gamma model[55] and others. Overall, the problem of rare events is well studied for the case of pairwise meta-analysis and many different models have been proposed to tackle the issue while those have also been evaluated through different simulation studies[42,43,56].

1.4. Basic concepts of network meta-analysis

Network meta-analysis (NMA) is a complex evidence synthesis tool that extends pairwise meta-analysis as it allows for the simultaneous comparison between any number of different competing interventions[57–59]. A network plot[60] is usually used to visualize the direct comparisons between the different interventions. The nodes in this plot represent the different competing interventions while the edges are linking those pairs of interventions which are directly compared

through at least one individual study. Different weighting schemes can be applied for the better representation of the data structure. For example, the size of the nodes in the network can be proportional to the number of participants receiving each intervention while the size of the edges can be proportional to the number of studies which compare the respective pair of interventions. **Figure 1** shows an example of a network which evaluates the safety of different interventions for chronic plaque psoriasis[61].

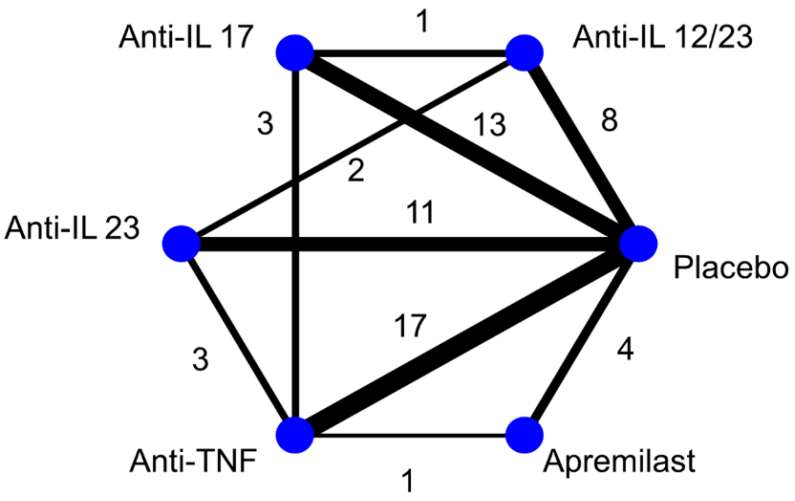


Figure 1: An example of a network plot. This network evaluates the safety of different interventions for chronic plaque psoriasis. The network is consisted by 5 classes of interventions and placebo. The edges in the plot are weighted according to the number of studies evaluating the respective comparison. Those numbers are also shown numerically in the plot.

NMA estimates are expected to be more precise compared to the summary intervention effects obtained from pairwise meta-analysis. This is because NMA combines both direct and indirect sources of information[15,62,63]. The direct comparisons arise from interventions being compared directly in individual studies while the indirect comparisons arise from indirect paths in the network. For example, for the network presented in **Figure 1** the class Anti-IL 23 can be compared with the class Anti-TNF both directly and indirectly. The direct evidence comes through

the three studies which compare those two classes. The indirect evidence arises through their comparison with their common comparator which is the Placebo intervention. In this case the resulting NMA estimate will be a mixed estimate that takes into account both direct and indirect estimates. In a more general setting, let a set of N studies to compare three different interventions namely 1, 2 and 3. An indirect estimate for the comparison between interventions 2 and 3 can be obtained by subtracting the direct estimates obtained for comparisons 3 vs 1 and 2 vs 1

$$\hat{\mu}_{23}^{ind} = \hat{\mu}_{13}^{dir} - \hat{\mu}_{12}^{dir}$$

A mixed estimate for the comparison 2 vs 3 can be obtained as a weighted average of the direct and indirect estimates as[62],

$$\hat{\mu}_{23}^{mixed} = \frac{\frac{1}{var(\hat{\mu}_{23}^{dir})}\hat{\mu}_{23}^{dir} + \frac{1}{var(\hat{\mu}_{23}^{ind})}\hat{\mu}_{23}^{ind}}{\frac{1}{var(\hat{\mu}_{23}^{dir})} + \frac{1}{var(\hat{\mu}_{23}^{ind})}}$$

where $var(\hat{\mu}_{23}^{ind})$ denotes the variances of the indirect estimate calculated as,

$$var(\hat{\mu}_{23}^{ind}) = var(\hat{\mu}_{13}^{dir}) + var(\hat{\mu}_{12}^{dir})$$

Finally, the variance of the mixed estimate is obtained as,

$$var\left(\hat{\mu}_{23}^{mixed}\right) = \frac{1}{\frac{1}{var(\hat{\mu}_{23}^{dir})} + \frac{1}{var(\hat{\mu}_{23}^{ind})}}$$

The notion of mixed estimates can be generalized in larger networks and allow for the incorporation of several indirect estimates through different paths in the network.

The validity of both indirect and mixed estimates depend on the validity of the core NMA assumption of transitivity[57]. In a network consisted of three interventions the transitivity assumption suggests that the common comparator intervention 1 is similar in the studies investigating the comparisons 3 vs 1 and 2 vs 1[63,64]. An alternative and more formal interpretation

of transitivity is that the missing arms across the available studies should be missing at random. This means that the choice of the interventions across the available studies should not be related with their effectiveness[65]. Transitivity can be assessed by comparing the distribution of the potential effect modifiers across the available set of studies[66]. If those distributions are not similar across two or more pairwise comparisons, then the transitivity assumption is very likely to be violated. However, due to the fact that the distribution of the different effect modifiers across the studies is very likely to be unknown, the transitivity assumption is often an untestable assumption[57].

The clinical and methodological differences across the studies that may lead to intransitivity can potentially be reflected in the data through disagreements between direct and indirect sources of evidence. The latter leads to the statistical manifestation of the transitivity assumption, namely the consistency assumption[57,65,67,68]. This assumption implies that the direct and indirect evidence in the network are in agreement regarding the true summary underlying intervention effects. Mathematically, in a network of three interventions the consistency assumption can be expressed using the consistency equations as,

$$\mu_{23} = \mu_{13} - \mu_{12} \tag{1.6}$$

The consistency assumption can be statistically evaluated as the parameters in equation (1.6) are estimable.

As already explained NMA models are extensions of pairwise meta-analysis models which can allow the simultaneous comparison between any number of competing interventions. Consider a network of N studies which assesses K different interventions. The aim of a NMA model is to estimate the relative effectiveness for all in total $\binom{K}{2}$ pairs of interventions. However, the

consistency equations in (1.6) suggests that only a set of $K-1$ basic comparisons is sufficient for the estimation of all the $\binom{K}{2}$ pairwise comparisons. All the rest, namely functional comparisons, can be estimated as linear combinations of the basic comparisons[57,59,63]. Thus, the total number of parameters which are needed to be estimated is $K-1$ denoted here as $\boldsymbol{\mu} = (\mu_{12}, \mu_{13}, \dots \mu_{1K})$. The basic comparisons are defined in terms of a common reference intervention which for this case is chosen to be the intervention 1 in the network. The choice of the reference group is generally arbitrary with the only restriction being that all interventions in the network should be expressed via at least one basic comparison.

Incorporating multi-arm studies is another advantage of NMA compared to pairwise meta-analysis which only accounts for two arm studies[69]. The inclusion of a K_i arm ($K_i \geq 2$) implies that the correlation existing for both the observed and the true study effects need to be taken into account. Under the consistency assumption a K_i arm study i provides intervention effects for in total $K_i - 1$ comparisons denoted here as $\boldsymbol{y_i} = (y_{i,12}, y_{i,13}, \dots, y_{i,1K_i})$. Thus, the within-studies variance covariance matrix of the observed study effects has the following form,

$$\mathbf{S_i} = \begin{pmatrix} s_{i,12}^2 & \cdots & cov(y_{i,12}, y_{i,1K_i}) \\ \vdots & \ddots & \vdots \\ cov(y_{i,1K_i}, y_{i,12}) & \cdots & s_{i,1K_i}^2 \end{pmatrix}$$

Moreover, the respective variance covariance matrix for true study effects $\boldsymbol{\theta_i} = (\theta_{i,12}, \theta_{i,13}, \dots, \theta_{i,1K_i})$ has the following form,

$$\mathbf{\Sigma_i} = \begin{pmatrix} \tau_{12}^2 & \cdots & cov(\theta_{i,12}, \theta_{i,1K_i}) \\ \vdots & \ddots & \vdots \\ cov(\theta_{i,1K_i}, \theta_{i,12}) & \cdots & \tau_{1K_i}^2 \end{pmatrix}$$

The off-diagonal elements in matrices $\boldsymbol{S_i}$ and $\boldsymbol{\Sigma_i}$ represent the covariance between two intervention effects (observed and true, respectively) in study i whether the diagonal elements represent the respective variances. The variances $\tau^2_{12}, \dots, \tau^2_{1K_i}$ represent the heterogeneity in the studies for each of the $K_i - 1$ pairwise comparisons. For simplicity, those paramerers across all comparisons in the network are usually assumed to be equal[70] (i.e. $\tau^2_{XY} = \tau^2 \; \forall X, Y \in \{1,2, \dots, K\}$). The assumption of a common heterogeneity parameter simplifies the calculation of the respective covariances as in this case those are equal to $\frac{\tau^2}{2}$. This means that under this assumption the variance covariance matrix of the true effects in study i has the following form,

$$\mathbf{\Sigma_i} = \begin{pmatrix} \tau^2 & \cdots & \frac{\tau^2}{2} \\ \vdots & \ddots & \vdots \\ \frac{\tau^2}{2} & \cdots & \tau^2 \end{pmatrix} \tag{1.7}$$

The structure of the variance covariance matrix given by equation (1.7) is the structure that will be assumed across the remaining parts of this thesis.

The hierarchical NMA model does not substantially differ from the corresponding approach for modeling a pairwise meta-analysis model. For a random-effects NMA model the equation (1.3) needs to replace the univariate normal distributions with the corresponding multivariate normal distributions. In this way a NMA model can be written as a hierarchical model in the following form[21,71],

$$\begin{aligned} \boldsymbol{y_i} &\sim N_{K_i-1}(\boldsymbol{\theta_i}, \mathbf{S_i}) \\ \boldsymbol{\theta_i} &\sim N_{K_i-1}(\boldsymbol{\mu_i}, \mathbf{\Sigma_i}) \end{aligned} \tag{1.8}$$

where $\boldsymbol{\mu_i}$ is the vector that includes the true intervention effects for all comparisons in study i. As in the case of pairwise meta-analysis the model in equation (1.8) can be fitted in both the frequentist and the Bayesian framework. In the last case prior distributions need to be assigned for all the K parameters of interest which are the $K-1$ basic comparisons and the heterogeneity parameter τ^2. By setting the heterogeneity parameter τ^2 equal to 0 the random-effects model in equation (1.8) renders to its common-effect counterpart. Alternative and equivalent representations of a NMA model may include the representations according to a meta-regression[72] model or a multivariate meta-analysis model[72–74] or a model arising from the electrical networks and graph theory[75].

Consistency or equivalently inconsistency checks are an integral part of NMA which can take place either locally or globally[57,65] in the network. The local approaches use as a measure of inconsistency the absolute difference between direct and indirect estimates, namely the inconsistency factor (IF). Based on the IF different methods have been proposed starting from the simplest one which is the loop-specific approach[76]. Suppose a network of C closed loops then the IF within each loop $c = 1,2,\dots,C$ can be calculated as,

$$\widehat{w}_{XY}^{c} = \left|\hat{\mu}_{XY}^{dir} - \hat{\mu}_{XY}^{C_c}\right| \tag{1.9}$$

with variance,

$$var(\widehat{w}_{XY}^{c}) = var\left(\hat{\mu}_{XY}^{dir}\right) + var\left(\hat{\mu}_{XY}^{C_c}\right) \tag{1.10}$$

for any comparison between interventions X and Y $(X, Y \in 1,2,\dots,K)$ in loop c. Then the loop-specific approach[76] investigates the null hypothesis $H_0: w_{XY}^{c} = 0$. The latter can be done using the following Z-statistic,

$$Z = \frac{w_{XY}^{c}}{\sqrt{var(w_{XY}^{c})}} \sim N(0,1) \quad (1.11)$$

The node-splitting and the back-calculation methods[67] are similar to the loop-specific approach. The only difference between those three methods lies in the way of calculating the indirect estimate. Specifically, the node-splitting approach excludes one direct comparison from the network each time. For example, suppose that all the direct evidence for the comparison of two interventions X and Y ($X, Y \in \{1,2, \dots, K\}$) is excluded from the network. Then the resulting NMA estimate for this comparison will be the indirect estimate $\hat{\mu}_{XY}^{ind}$as all the direct evidence was previously excluded from the analysis. Based on equations (1.9), (1.10) and (1.11) one can estimate the statistical significance of the inconsistency factor $\widehat{w}_{XY}^{c}$ between each pair of interventions in the network. For the back-calculation method the indirect estimate can be obtained as,

$$\hat{\mu}_{XY}^{ind} = \left(\frac{\hat{\mu}_{XY}^{mixed}}{var(\hat{\mu}_{XY}^{mixed})} - \frac{\hat{\mu}_{XY}^{dir}}{var(\hat{\mu}_{XY}^{dir})}\right) var(\hat{\mu}_{XY}^{ind}), X, Y \in \{1,2,..,K\}$$

In this case the variances of the above estimates are connected through the equation,

$$\frac{1}{var(\hat{\mu}_{XY}^{ind})} = \frac{1}{var(\hat{\mu}_{XY}^{mixed})} - \frac{1}{var(\hat{\mu}_{XY}^{dir})}$$

Again, based on equations (1.9), (1.10) and (1.11) one can estimate the statistical significance of the inconsistency factor $\widehat{w}_{XY}^{c}$ between each pair of interventions in the network.

For Bayesian NMA models the aforementioned methods can be applied with the presence or not of inconsistency being indicated from the comparison between the posterior densities of direct and indirect estimates for every comparison X and Y. Consistency checks are a crucial step from every NMA, thus many other alternative and more sophisticated methods have been

proposed. The composite test for inconsistency[77] extends the loop-specific approach to allow for discrepancies between several indirect estimates in the network. Finally, methods for globally checking for inconsistency can include either global tests for inconsistency or global models. Those models are known as inconsistency models such and well-known examples are the unrelated mean effects model[78], the Lu-Ades[65] model and the design by treatment interaction model[79–81].

One of the most attractive properties of NMA is that all the evidence coming from a network of interventions can be combined and result in the final relative ranking for the different competing interventions[45]. Different measures have been proposed for calculating the relative ranking of the different interventions in a network. Those measures are usually based on the ranking probabilities π_{tk} which are interpreted as the probability that the k^{th} intervention in the network will be at rank t $(k, t \in \{1,2, \dots K\}$. A popular ranking metric in the frequentist setting[72,82,83] is the P-score method[84] which is based on the multivariate normal distributions with mean and variance arising from the NMA estimates. A P-score can be calculated as

$$\mathrm{P}_k = \frac{1}{K-1} \sum_{t, t \neq k}^{K} p_{k>t}$$

where $p_{k>t}$ denotes the probability that the outcome of intervention k is better than that of intervention t. The graphical representation of the P-scores can take place using the rankograms[83].

In the Bayesian framework ranking measures arise from a large sample obtained by the posterior distribution of the final NMA estimates. Examples of Bayesian ranking metrics can be the mean rank[85] or the the method based on the surface under the cumulative ranking area[83] (SUCRA). The mean rank metric (mrank) can be calculated as,

$$\text{mrank}_k = \sum_{t=1}^{K} \pi_{tk} * t$$

and the SUCRA metric can be calculated as,

$$\text{SUCRA}_k = \frac{1}{K-1} \sum_{t=1}^{K-1} \sum_{x=1}^{t} \pi_{tx}$$

Those two Bayesian metrics are equivalent as it can be shown the $\text{SUCRA}_k = \frac{K - \text{mrank}_k}{K-1}$. The SUCRA for intervention k can be interpreted as the average proportion of interventions in the network which are worse than k.

1.5. The problem of sparse data in network meta-analysis

Problems with sparse data can still occur at the highest level of evidence which comes from NMA models. This is expected as the IV NMA model given in equation (1.8) has the same distributional assumptions with the pairwise meta-analysis model given in equation (1.3). The case of sparse data in terms of small sample sizes appears for NMA in the form of the sparse networks. Those are cases of networks with limited number of direct comparisons and only few studies to inform them. The case of rare events can appear similarly to the form that it appears for pairwise meta-analysis as the case where the binary outcome of interest is rarely observed across the available studies.

1.5.1. Conducting network meta-analysis in presence of sparse networks

Results obtained from sparse networks are expected to be accompanied with substantial uncertainty not only in the estimation of the intervention effects but also in the validity of the underlying NMA assumptions[29,86,87]. The transitivity assumption requires the existence of balanced distributions for the effect modifiers across the available studies. Transitivity checks can be very challenging in the presence of sparse networks as it can be expected that such distributions can materially differ if only a small set of studies is considered[86]. Moreover, the formal methods for checking the consistency assumption involve statistical tests for checking the potential discrepancies between direct and indirect estimates. In cases of sparse networks those tests can be substantially underpowered and thus fail to properly identify such discrepancies [45,67]. Additionally, consistency checks can be challenging in the Bayesian setting as the posterior densities of both direct and indirect estimates are expected to be very wide[67]. This can mask important differences across the two posterior densities and thus potential inconsistencies in the network can pass unnoticed. Overall, sparse networks can challenge the formal evaluation of both transitivity and consistency assumptions and thus they can threaten the validity of the final NMA estimates.

The distributional assumption made by the IV NMA model can be problematic in the presence of sparse networks. In such cases the approximate normal assumptions are invalid thus leading to biased and imprecise estimates for all the $K-1$ basic comparisons in the network. As explained in section 1.3.1 in presence of few available studies the estimation of the heterogeneity parameter is problematic[34,35]. Biased estimates of the heterogeneity parameter can lead to improper confidence or credible intervals for the final NMA estimates and finally to unreliable conclusions about the intervention effectiveness or safety[88]. Recent publications aim to deal with the problem

of heterogeneity estimation in NMA of sparse networks by allowing the incorporation of external evidence which can increase the amount of the available information and facilitate the estimation of this parameter[87,88]. In addition, solutions like the use of empirical prior distributions[33,89,90] or the use of weakly informative prior distributions[53,91] can be borrowed from Bayesian pairwise meta-analysis to potentially facilitate the heterogeneity estimation when conducting a Bayesian NMA.

The presence of sparse networks suggests a very important issue for NMA.This can become even more serious considering that it can frequently occur. A recent empirical study found 92 (7,4%) published NMAs that included more interventions than the number of studies in a sample of 1236 NMAs of RCTs with at least 4 interventions[92]. On top of that, in a review of 186 NMAs it was found that the median number of studies per network was 21 and the median number of studies per comparison was 2[93]. Overall, the estimation of the intervention relative effectiveness in presence of sparse networks can be very much challenging and the final NMA estimates should always be interpreted with cautiousness.

1.5.2.Conducting network meta-analysis in presence of rare events

Similarly, to pairwise meta-analysis the problem of rare events can seriously threaten the validity of the findings of a NMA model. NMA with rare events has attracted little attention in the published literature compared to the attention that has been given to the problem for pairwise meta-analysis[41–43]. Until recently networks of interventions with rare events were analyzed only based on the IV NMA model. However, as described previously the assumptions made by this model can be invalid in presence of rare outcomes and as a consequence the NMA estimates are expected to be biased and imprecise.

An additional challenge in NMA lies in handling studies that report zero events in one or all interventions arms. The IV model can allow the inclusion of those studies by using a continuity correction for studies with zeros. However, such corrections have previously been found to introduce bias in the final estimates[42,43]. The exclusion of those studies from the analysis is usually suggested for the case of pairwise meta-analysis[14]. The latter can be also adopted for the case of NMA but in this case one should also consider the additional issue of network connectivity. The extensive exclusion of studies with zeros may cause a dense network of interventions with rare events to become sparse in terms of the available studies for analysis. Additionally, such exclusions may lead to the extreme case where initially connected networks become disconnected and thus in such cases even conducting a NMA can be pointless.

The problematic distributional assumptions of the IV NMA model can be replaced using the non-parametric extension of the MH method for a NMA model[94]. Even though the MH for pairwise meta-analysis can be used for various effect measures such as OR, RR or RD for NMA the respective extension refers only to the calculation of the OR[94]. The method was compared in a simulation study with the IV model and the frequentist common-effect non-central hypergeometric (NCH) model[54]. This simulation showed that the MH method can be a reliable alternative for NMA with rare events. Additionally, the IV model was found to be unsuitable as in most cases it provided the most biased results. In terms of heterogeneity estimation the DL method was found to be biased and underpowered as it was resulting in $\hat{\tau}^2 = 0$ across scenarios which assumed a $\tau^2 \neq 0$. Alternative methods can include the use of the exact binomial distribution or the NMA models fitted in the Bayesian framework. However, those choices are never evaluated in simulations for NMA with rare events[41,54,94]. Finally, the non-parametric Peto method has been found unsuitable

to be used for NMA as this method fails to properly account for multi-arm trials and thus to give valid indirect estimates[95].

1.6. Aims and Objectives of the thesis

The aim of this thesis is to provide suitable frameworks for analyzing sparse data in the meta-analytical field. The main objective is through those frameworks to facilitate the reliable estimation of the intervention effects in both cases of rare events and few available studies. A second objective of the thesis is to compare the proposed frameworks with other existing methodologies and thus to provide final recommendations. Finally, this thesis aims to enable the broad use of the proposed methodologies through user-friendly environments which everyone can freely access and use without having any sophisticated programming skills.

Three different projects were conducted to deal with sparse data. The first project focused on the issue of rare events in NMA. In this project a new framework for conducting NMA with rare events is proposed. This framework arises as a generalization of penalized likelihood method which is widely used for the analysis of rare events in individual studies but it was never extended in the field of NMA. An extensive simulation study was conducted to evaluate the performance of the proposed method and to compare it with other existing methodologies.

The second project of this thesis aimed to allow for the broad use of the methods proposed in the first project and to deal with the issue of rare events in the context of Covid-19 trials. This appeared to be very important as in the context of meta-analysis of Covid-19 trials the occurrence of rare events was frequent and thus the use of the IV model should be limited in such circumstances. Specifically, in this project an R-Shiny application was constructed to allow for the

methods proposed to deal with rare events to be used through a user-friendly environment that everyone can use without requiring any programming skills.

Finally, the third project of this thesis deals with the issue of sparse networks. In this project it a framework was developed for sharing information between a sparse and a dense network of interventions which pertain into different subgroups of the population. The proposed framework was compared with other potential solutions and the standard NMA model through an illustrative real example of two networks which investigate the effectiveness of different antipsychotics.

2. Part 1: Dealing with rare events in network meta-analysis

Rare events can cause important problems in the validity of the statistical analysis across various fields of evidence based medicine[1,2,41]. As already explained in Chapter 1 the problems can start even from the phase of individual studies and they can extend to higher levels of evidence such as pairwise meta-analysis or NMA. In NMA, the problems arise from the approximate normal assumptions which are used by the standard NMA model which is the IV model[94]. The assumption of fixed within-study variances can be more pragmatic in presence of large individual studies with an adequate number of observed events. In this case the large sample approximations can lead to robust estimates which are close to the true value[45]. However, in the presence of rare events those large sample approximations are expected to be invalid and thus the study-specific estimates of variance are usually biased. A second issue arises from the failure of the different methods to properly estimate the heterogeneity parameter in presence of rare events. In such cases the most of the methods underestimate τ as 0 thus rendering a random-effects NMA model to its common-effect counterpart[41,46]. Overall, the problems related to the distributional assumptions of the IV NMA model are expected to lead to biased NMA estimates for estimating the relative interventions effects.

A potential way to tackle rare events can arise from the replacement of the problematic within-studies normal assumption with the exact binomial distribution[21,51]. In this way a NMA model can be fitted by using logistic regression. This is a one-stage NMA model which models the events r_{ik} observed in arm $k = 1,2, \dots, K$ of study $i = 1,2, \dots, N$ using a binomial distribution $Bin(n_{ik}, p_{ik})$ where n_{ik} and p_{ik} denote the sample size and the probability of event in arm k of study i. In this

way the one-stage pairwise meta-analysis model given in equation (1.5) can be extended with only minor modifications to a NMA model which is parametrized as[51],

$$logit(p_{ik}) = \alpha_i + \delta_{i,b_{(i)}k}, i = 1,2, \dots N, k = 1,2, \dots, K \quad (2.1)$$

$$\delta_{i,b_{(i)}k} \sim N(d_{b_{(i)}k}, \tau^2)$$

where α_i is the fixed study-specific intercept interpreted as the log-odds of event in the baseline group and $\delta_{i,b_{(i)}k}$ is the study-specific true logOR for the comparison of intervention $k \in \{1,2, \dots, K\}$ versus the reference intervention $b_{(i)} \in \{1,2, \dots, K\}$ in study i. The random-effects $\delta_{i,b_{(i)}k}$ are assumed to follow a normal distribution with grand mean equal to $d_{b_{(i)}k}$ and variance τ^2 to represent the across-studies heterogeneity. Obviously, $\delta_{i,b_{(i)}k} = 0$ when $b_{(i)} = k$. The model in (2.1) can be reduced to a common-effect model if the parameter τ^2 is set to 0. In this case the parameter $\delta_{i,b_{(i)}k}$ in (2.1) needs to be replaced with $d_{b_{(i)}k}$.

The model in (2.1) relies on the maximum likelihood theory for estimating the unknown parameters. These MLEs are known to be biased in cases where the large sample approximations are invalid[8]. The latter is also known for the case of logistic regression models where rare events can lead to underestimation of the probabilities of event $\hat{p}_{ik}$ and thus overestimation of the intervention effects[10]. Moreover, the issue of handling studies with zero events in one or all intervention arms can also appear for model (2.1) as this problem resembles the problem of separation[53]. In cases of separated data, the MLE is known to be infinite with an infinite variance. Finally, it should be noted that the estimation of the heterogeneity parameter in model (2.1) relies on either the maximum likelihood or the REML method thus the estimation of this parameter can also be problematic in cases of rare events. Specifically, it has been found for pairwise meta-

analysis that the estimates of τ^2 either model (2.1) or equivalently (1.5) are usually biased with τ^2 being underestimated as 0[46] thus indicating no heterogeneity across the studies.

The bias of the MLE has been studied extensively in the literature for individual studies. A common characteristic across the proposed approaches is that they target the first term $\frac{b_1(\boldsymbol{\beta})}{n}$ of the asymptotic bias expansion in (1.1) and aim to remove it from the MLE[8]. This happens as this term carries the biggest amount of bias which is incorporated in the estimate $\widehat{\boldsymbol{\beta}}$[8]. Some well known methods for correcting the bias of the MLE are the jackknife[96] and the bootstrap[97] methods. Those correct the bias through a two step procedure[98,99]. At the first step the bias expansion $B(\boldsymbol{\beta})$ in (1.1) is estimated and at the second step the new less biased MLE is derived as $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \hat{B}(\boldsymbol{\beta})$. However, those methods have the drawback of explicitly relying on the initial value $\widehat{\boldsymbol{\beta}}$. This can become more evident in presence of separation phenomena as in such cases the difference $\widehat{\boldsymbol{\beta}} - \hat{B}(\boldsymbol{\beta})$ remains undefined and the bias correction methods cannot be applied.

An alternative for individual studies can be the use of the penalized likelihood regression models. Specifically, a penalization to the likelihood function was proposed by Firth[100] in order to improve the performance of the MLE in terms of bias and to provide MLEs which are free of the first term of the asymptotic bias expansion in (1.1). Firth showed that the penalization of the likelihood function with Jeffrey's invariant prior[101] can result in MLEs which are less biased than the non-penalized estimates. Beyond the nice properties in terms of bias reduction, the penalized estimates can also solve separation issues. Specifically, Heinze and Schemper[102] showed through a simulation study that the penalization with Jeffrey's invariant prior can yield to finite estimates even in presence of separation. Finally, recently Kosmidis and Firth[103] mathematically proved the

findings of Heinze and Schemper by showing that the penalized likelihood function is strictly concave and thus the penalized likelihood estimate is always finite.

Even though the penalized likelihood method proposed by Firth is a common method for analyzing rare events in individual studies the method has never been extended and studied for the case of NMA with rare events. Given that, the aim of this project was to adopt and generalize this method for the field of NMA. To do so, a common-effect NMA model which uses logistic regression was at first adopted and its likelihood function was penalized with Jeffrey's invariant prior. This model constitutes the proposed methodology of a penalized likelihood network meta-analysis model (PL-NMA). The PL-NMA model can preserve the connectivity of the network as the with the penalization approach there is no need to exclude studies reporting zero events. This model was initially proposed as a common-effect model and it was extended to a random-effects model by treating heterogeneity as an overdispersion parameter. This means that the heterogeneity was added for the PL-NMA model as a multiplicative parameter φ[104,105]. Specifically, a two-stage approach was adopted were at the first-stage the common-effect PL-NMA model is fitted and at the second-stage the variances of the resulting NMA estimates are inflated based on their multiplication with the parameter φ.

Both the common and the random-effect(s) models were evaluated through an extensive simulation study of in total 33 scenarios. The PL-NMA model was compared in this simulation with the IV NMA model, the MH NMA model, the random-effects logistic regression model given by equation (2.1) and finally three logistic regression models fitted in the Bayesian framework. Those models were compared in terms of various performance measures such as the mean bias or the coverage probability.

The PL-NMA method was found in the simulation to outperform the rest of the approaches in terms of bias while the IV NMA model was once again found to be a suboptimal choice of model for analyzing rare events. Additionally, the impact of studies with zero events was also assessed in the simulation study. The PL-NMA model was fitted once by including and once by excluding zero event studies and the performance of the method was evaluated in both of those circumstances. Overall, both the inclusion and the exclusion approaches appeared to have a small impact in the estimation of the OR.

The detailed description of the method and the findings of this work can be found in the following open-access article: Evrenoglou T, White IR, Afach S, Mavridis D, Chaimani A. Network meta-analysis of rare events using penalized likelihood regression. Statistics in Medicine. 2022;1-17. doi: 10.1002/sim.9562.

The online supplementary files of the manuscript can be found in Annex 1 of this thesis.

Received: 3 September 2021 | Revised: 28 July 2022 | Accepted: 12 August 2022

DOI: 10.1002/sim.9562

RESEARCH ARTICLE

Statistics in Medicine WILEY

Network meta-analysis of rare events using penalized likelihood regression

Theodoros Evrenoglou[1] | **Ian R. White**[2] | **Sivem Afach**[3] | **Dimitris Mavridis**[1,4] | **Anna Chaimani**[1,5]

[1]Université Paris Cité, Research Center of Epidemiology and Statistics (CRESS-U1153), INSERM, Paris, France

[2]MRC Clinical Trials Unit, University College London, London, UK

[3]Université Paris-Est Créteil, UPEC, Créteil, France

[4]Department of Primary Education, University of Ioannina, Ioannina, Greece

[5]Cochrane France, Paris, France

Correspondence
Theodoros Evrenoglou, Université Paris Cité, Research Center of Epidemiology and Statistics (CRESS-U1153), INSERM, Paris, France, Hôpital Hôtel-Dieu, 1 Place du Parvis Notre-Dame, 75004 Paris, France.
Email: tevrenoglou@gmail.com

Funding information
Medical Research Council, Grant/Award Number: MC_UU_00004/06; European Union's Horizon 2020 COMPAR-EU project, Grant/Award Number: No754936

Network meta-analysis (NMA) of rare events has attracted little attention in the literature. Until recently, networks of interventions with rare events were analyzed using the inverse-variance NMA approach. However, when events are rare the normal approximations made by this model can be poor and effect estimates are potentially biased. Other methods for the synthesis of such data are the recent extension of the Mantel-Haenszel approach to NMA or the use of the noncentral hypergeometric distribution. In this article, we suggest a new common-effect NMA approach that can be applied even in networks of interventions with extremely low or even zero number of events without requiring study exclusion or arbitrary imputations. Our method is based on the implementation of the penalized likelihood function proposed by Firth for bias reduction of the maximum likelihood estimate to the logistic expression of the NMA model. A limitation of our method is that heterogeneity cannot be taken into account as an additive parameter as in most meta-analytical models. However, we account for heterogeneity by incorporating a multiplicative overdispersion term using a two-stage approach. We show through simulation that our method performs consistently well across all tested scenarios and most often results in smaller bias than other available methods. We also illustrate the use of our method through two clinical examples. We conclude that our "penalized likelihood NMA" approach is promising for the analysis of binary outcomes with rare events especially for networks with very few studies per comparison and very low control group risks.

KEYWORDS
bias reduction, maximum likelihood estimates, multiple treatment meta-analysis, rare endpoints

1 | INTRODUCTION

Network meta-analysis (NMA) has become an essential tool of comparative effectiveness research.[1] Despite the rapid and extensive development of NMA methods over the last decade,[2] little attention has been given to the issue of rare events within a network of interventions. Typically, NMAs with rare events are performed using the standard "inverse-variance" (IV) model with a continuity correction for studies with zero events. This is a two-stage contrast-based model where

at the first-stage the effect sizes and their variances are extracted from the studies and at the second-stage these are synthesized to obtain the summary treatment effect estimates. A major drawback of this approach is that it relies on large sample approximations and normality assumptions which are implausible under the presence of low number of observed events. An alternative popular approach, usually fitted in Bayesian framework, is based on the exact binomial distribution. However, it has been suggested that Bayesian methods may be problematic for meta-analyzes of rare events since results can be dominated by the prior distribution.[3,4] On the other hand, when such a model is fitted in frequentist framework it relies on the maximum likelihood estimates (MLE) which are known to be biased in the presence of rare events.[5,6] Other, possibly less biased, options for performing NMA with rare events are the recent extension of the Mantel-Haenszel (MH) meta-analytical model[4] and a model based on the noncentral hypergeometric distribution (NCH).[3] The MH model, though, is a nonparametric common-effect model and, thus, it avoids the use of any distributional assumption.

A further important issue of most meta-analyzes with rare events is how to handle studies with zero events in all treatment groups. Imputations of arbitrarily chosen constants (eg, 0.5) have been found to bias the results.[7,8] More sophisticated models (eg, random intercepts models) allow inclusion of such studies by using between-study information which, however, may again bias the results.[9,10] Existing methods for NMA with rare events that avoid the use of between-study information, like the aforementioned MH-NMA and NCH-NMA models, either exclude these studies from the analysis or require continuity corrections. Studies with zero events in all treatment groups are not infrequent in meta-analyzes of rare endpoints and the optimal way to treat such studies is still unclear. A recent meta-epidemiological study of 442 Cochrane reviews including at least one study with no events in both treatment groups suggested that the inclusion or exclusion of these studies can impact materially the results of the meta-analyses.[11] In addition, in the context of NMA, exclusion of these studies may lead to disconnected networks.

In the present article, we aim to tackle the problem of rare events in networks of interventions with binary data by adapting and extending methodology from the analysis of individual studies. We use the logistic regression expression of NMA[12] and the penalization to the likelihood function proposed by Firth[13] to estimate the NMA relative effects, using odds ratios, through an one-stage model. In this way, we attempt to likely obtain less biased and more precise estimates in comparison to the existing methods for NMAs with rare events described earlier. On top of that, our penalized likelihood NMA method (PL-NMA) allows the inclusion of studies with zero events in all treatment groups without using between-study information. The rest of the article is structured as follows. In Section 2, we describe the standard NMA model as a logistic regression model and the proposed PL-NMA model. Section 3 presents a simulation study exploring the performance of our method in terms of bias and precision in comparison to existing NMA approaches for rare events under different scenarios. Finally, in Section 4 we apply the different models in two exemplar NMA datasets, and in Section 5 we discuss the strengths and limitations of our PL-NMA approach.

2 | METHODS

2.1 | Common-effect NMA using logistic regression

Suppose a network of N studies and T treatments. Let r_{ik} and n_{ik} denote the number of events and the number of total participants, respectively, in treatment group k of study i where $i = 1, 2, \dots, N$ and $k \in K_i$ with $K_i = \{\textit{treatments evaluated in study } i\}$. We assume that

$$r_{ik} \sim Bin\left(n_{ik}, p_{ik}\right),$$

where p_{ik} is the probability of an event in treatment group k of study i. Then, we model the probabilities p_{ik} through the following logistic regression model,

$$logit\left(p_{ik}\right) = \alpha_i + d_{b_i k} \tag{1}$$

with b_i an arbitrarily chosen baseline treatment from the set K_i.

Parameter $d_{b_i k}$ represents the log-odds ratio ($logOR$) of treatment k vs b_i in the ith study with $d_{b_i b_i} = 0$ for $k = b_i$. The parameter α_i represents the log-odds of the event in the b_i group and it is treated as a nuisance parameter. Under the

transitivity assumption, the *logOR* between any two treatments t_1 and t_2 ($t_1, t_2 = 1, \ldots, T$ and $t_1 \neq t_2$) is

$$d_{t_1 t_2} = d_{b_i\, t_2} - d_{b_i\, t_1}.$$

It follows from (1) that,

$$p_{ik} = expit\left(\alpha_i + d_{b_i k}\right).$$

Let $\boldsymbol{\alpha}$ denote the vector that contains all study intercepts α_i, $\boldsymbol{d}$ the vector of the relative effects $d_{b_i k}$, and $\boldsymbol{r}$ and $\boldsymbol{n}$ the vectors of observed events and sample sizes per study and treatment arm.

The estimation of the parameters $d_{b_i k}$ relies on the maximization of the likelihood function which can be written as,

$$L(\boldsymbol{\alpha}, \boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{n}) = \prod_{i=1}^{N} \prod_{k \in K_i} \binom{n_{ik}}{r_{ik}} expit\left(\alpha_i + d_{b_i k}\right)^{r_{ik}} \left(1 - expit\left(\alpha_i + d_{b_i k}\right)\right)^{n_{ik} - r_{ik}}.$$

Equivalently, the log-likelihood function is,

$$\mathrm{l}(\boldsymbol{\alpha}, \boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{n}) = \sum_{i=1}^{N} \sum_{k \in K_i} log \binom{n_{ik}}{r_{ik}} + r_{ik} \log\left(expit\left(\alpha_i + d_{b_i k}\right)\right) + (n_{ik} - r_{ik}) log\left(1 - \mathrm{expit}\left(\alpha_i + d_{b_i k}\right)\right). \tag{2}$$

The above model relies on the maximization of the log-likelihood in Equation (2) in terms of the $N + T - 1$ parameters $\boldsymbol{\alpha}$ and $\boldsymbol{d}$. These MLE are biased in the case of finite sample sizes and their unbiasedness can only be assumed approximately for a large sample size and a considerable number of observed events.[5] Hence, the MLE can be problematic when the number of observed events is small, underestimating the probability of an event and overestimating the treatment effects.[14,15]

2.2 | Common-effect penalized likelihood NMA

To reduce the bias of the above MLE for NMAs with rare events, we employ Firth's modification[13] to the log-likelihood function in (2); this is the frequentist equivalent to using as penalty Jeffrey's invariant prior. This modification results in the penalized likelihood and log-likelihood functions, respectively,

$$L^*(\boldsymbol{\alpha}, \boldsymbol{d} | \boldsymbol{r}, \boldsymbol{n}) = L(\boldsymbol{\alpha}, \boldsymbol{d} | \boldsymbol{r}, \boldsymbol{n})\ |\boldsymbol{I}|^{\frac{1}{2}}$$

and

$$l^*(\boldsymbol{\alpha}, \boldsymbol{d} | \boldsymbol{r}, \boldsymbol{n}) = l(\boldsymbol{\alpha}, \boldsymbol{d} | \boldsymbol{r}, \boldsymbol{n}) + \frac{1}{2} log|\boldsymbol{I}|. \tag{3}$$

The matrix $\mathbf{I} \equiv \mathbf{I}(\boldsymbol{\alpha}, \boldsymbol{d})$ is the Fisher's information matrix with dimensions $(N + T - 1) \times (N + T - 1)$ that is given as

$$\boldsymbol{I} = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} \tag{4}$$

with $\mathbf{Z}$ being the model's design matrix with dimensions $\sum_{i=1}^{N} A_i \times (N + T - 1)$ and entries 1 for the columns associated with the relevant treatment and study and 0 otherwise.[16] A_i is the number of arms in study i, $\mathbf{W} = \mathbf{diag}\ \{ n_{ik} p_{ik} (1 - p_{ik}) \}$ and $|\mathbf{I}|$ is the determinant of $\mathbf{I}$. The penalized likelihood function is being maximized with respect to the $(N + T - 1)$ parameters $\boldsymbol{\alpha}$ and $\boldsymbol{d}$.

Equation (3) arises by the Taylor series expansion of the score function (ie, derivative of the log-likelihood function). Firth showed that the use of Jeffrey's invariant prior in this expansion penalizes the likelihood function and provides less biased estimates than the standard likelihood[13] function of Equation (2). The estimates obtained from the penalized likelihood function are typically shrunk toward 0 and this in turn results in smaller SEs. This is why the estimates obtained from the penalized likelihood function are expected to be more precise than those obtained from the standard likelihood

function.[17] The above PL-NMA provides finite treatment effects and SEs even in the case of studies that report zero events in all treatment arms.[17,18] Specifically, the issue of studies reporting only zero events in meta-analysis resembles that of separation (ie, when one or more covariates perfectly predict the outcome) in individual studies.[19,20] The PL-NMA method allows the inclusion of studies with zero events in two or more groups, without sharing information between the studies, by adding some prior information to the probability of an event through Jeffrey's prior. The latter usually results in estimated probabilities of event which are shifted toward 0.5.[17,20]

After maximizing the penalized likelihood function, we can replace p_{ik} with $\widehat{p}_{ik}$ in matrix $\mathbf{W}$. Those probabilities can be easily obtained as $\widehat{p}_{ik} = expit\left(\widehat{a}_i + \widehat{d}_{b_ik}\right)$. The square roots of the diagonal elements of the $\mathbf{I^{-1}}$ represent the SEs of the estimates which are used to construct Wald type confidence intervals; for instance, the square root of the diagonal element in the first row and first column of matrix $\mathbf{I^{-1}}$ represents the SE of the parameter in the first column of matrix $\mathbf{Z}$. The matrix $\mathbf{I^{-1}}$ is calculated as the inverse of matrix $\mathbf{I}$ in Equation (4). An alternative option is to construct profile likelihood confidence intervals that can be obtained using the critical region defined by the inequality

$$2\left\{l^*(\widehat{\alpha}, \widehat{\boldsymbol{d}}|\boldsymbol{r}, \boldsymbol{n}) - l^*\left(\widehat{\alpha}, \widehat{\boldsymbol{d}}_{\boldsymbol{0}}|\boldsymbol{r}, \boldsymbol{n}\right)\right\} \leq c_{1,1-q},$$

where $\widehat{\boldsymbol{d}}_{\boldsymbol{0}} = \left(d_{b_ik_0}, \widehat{\boldsymbol{d}}_{\boldsymbol{b_ik}}\right)$ is the vector that contains the estimated treatment effects of all treatments in the network vs the reference, except $k_0 \neq k$ which is a specific treatment. The boundary $c_{1,1-q}$ is the $(1-q)$ quantile of the chi-square distribution with 1 degree of freedom.

In case of $T = 2$, the model in (1) with the likelihood function in (3) corresponds to the penalized likelihood regression model for pairwise meta-analysis.

2.3 | Random-effects penalized likelihood NMA

The model in Equation (1) can be extended to the so-called binomial-normal (BN-NMA) model if we replace the d_{b_ik} with δ_{i,b_ik} which are now the study-specific treatment effects that are assumed to follow a normal distribution with mean d_{b_ik} and variance τ^2. In this model, the true treatment effects vary from study to study but the study intercepts α_i are kept fixed. The model can also allow for random intercept terms. The random-intercepts model requires a common reference treatment across the studies.[21] This theoretical common reference does not have to be observed within each study but its impact needs to be independent from unmeasured covariates. The PL-NMA model is a common-effect model and its extension to incorporate heterogeneity is challenging. In particular, the standard approach of incorporating heterogeneity as an additive term is not straightforward as the penalization requires closed forms of the likelihood function and its moments[13] and these are not available for a logistic regression mixed-effect model. Therefore, we use an alternative approach previously suggested in the literature where heterogeneity is incorporated as a multiplicative term.[22-24] This is a two-stage approach that modifies the study variances through an overdispersion parameter. Specifically, after fitting the model presented in the previous section, we multiply the variance in study i and arm k $v_{ik} = n_{ik}p_{ik}(1-p_{ik})$ with a scale parameter φ,[21-23] hence

$$v^*_{ik} = v_{ik} * \varphi, \quad \varphi \geq 1.$$

This implies that the matrix $\mathbf{W}$ in Equation (4) is replaced by $\varphi\mathbf{W}$. In this way, the point estimates of the treatment effects remain unaffected but their variances are inflated by that unknown parameter φ. To estimate the parameter φ we use the expression suggested by Fletcher[25]

$$\widehat{\varphi} = \frac{\widehat{\varphi}_P}{1+\bar{s}},$$

where $\bar{s}$ is the mean value of $s_{ik} = \left(\frac{\partial\widehat{v}_{ik}}{\partial\widehat{p}_{ik}}\right)\frac{\left(r_{ik}-\widehat{E}(r_{ik})\right)}{\widehat{v}_{ik}}$, $\frac{\partial\widehat{v}_{ik}}{\partial\widehat{p}_{ik}}$ is the derivative of $\widehat{v}_{ik}$ with respect to $\widehat{p}_{ik}$ and $\widehat{\varphi}_P = \frac{P}{m}$ with P denoting the Pearson's statistic, and $m = (T-1)+(N-1)$ the residual degrees of freedom of the model. The Pearson's

statistic for the NMA model presented in the previous section is,

$$P = \sum_{i=1}^{N} \sum_{k \in K_i} \frac{r_{ik} - \widehat{E}(r_{ik})}{\widehat{v}_{ik}},$$

where $\widehat{E}(r_{ik}) = n_{ik} * \widehat{p}_{ik}$. If $\widehat{\varphi} < 1$, we set it equal to 1.

3 | SIMULATION

We conducted a simulation study to compare our proposed and existing methodologies for NMA of rare events. We report our simulation following the recommendations by Morris et al[26]

3.1 | Data generation

We start the data generation process by specifying the number of treatments in the network (3, 5, or 8) and the number of studies in every treatment comparison (2, 4, or 8). We always create all possible $\frac{T(T-1)}{2}$ treatment comparisons; hence we only assume fully-connected networks. We then generate the total number of participants per study arm using a uniform distribution, namely, $n_{ik} \sim Unif(c_1, c_2)$, with $c_1 = 30$, $c_2 = 60$ for generating small studies and $c_1 = 100$, $c_2 = 200$ for larger studies. We only generated studies with an equal number of participants across arms. For each study we generated the control group risk; that is the risk of the event for patients receiving the reference treatment irrespective of whether this was evaluated in the study, thus for the simulation $b_i \equiv 1$. We generate the risk of an event in the reference treatment group 1 using $p_{i1} \sim Unif(u_1, u_2)$, with $u_1 \in [0.5\% - 5\%]$, $u_2 \in [1\% - 10\%]$ depending on the scenario (see Table 1). In all scenarios, the true *logORs* were fixed at equal intervals between 0 and 1 (eg, in a network of 5 treatments the true *logORs* are set to 0.25, 0.5, 0.75, and 1 for the comparisons of treatments 2, 3, 4, and 5 vs the reference treatment 1, respectively) and were used to calculate the odds of an event in every study arm,

$$odds_{i1} = \frac{p_{i1}}{1 - p_{i1}},$$

$$odds_{ik} = odds_{i1} * OR, k \neq 1.$$

Finally, we obtain the event probabilities in the non-reference treatment groups as $p_{ik} = \frac{odds_{ik}}{1+odds_{ik}}$ and the number of events in every study arm using a binomial distribution $r_{ik} \sim Bin(n_{ik}, p_{ik}), \forall i, k$.

For scenarios assuming the presence of heterogeneity ($\tau = 0.1$), we set a common parameter τ to represent the SD of the random-effects (see the Supplementary material S1 for a description).

We explored in total 33 scenarios with varying control group risks in the studies, number of treatments in the network, study size, number of studies per comparison, and magnitude of heterogeneity (Table 1). In scenarios 1–16 we considered only two-arm studies while in scenarios 17–32 we only included multi-arm studies evaluating all treatments. The latter are certainly not very realistic scenarios but they aimed at exploring whether the increased correlation from multi-arm studies could contribute to precision and bias reduction in NMAs with rare events. In scenario 33, we set extremely low control group risks (Table 1) to increase the number of generated studies with zero events in all treatment arms. This scenario only included two-arm studies.

We considered both homogeneous and heterogeneous cases and with different number of studies per comparison. Scenarios 1–10 only considered small studies (ie, 30–60 participants per arm). For scenarios 11–32, we increased the number of patients per arm and we mostly focus on cases in which the control group risks were extremely low (ie, 1%–2% and 0.5%–1%). In Scenarios 21–24 and 29–32 we considered a mix of low with higher control group risks. Table 1 provides a summary of the different scenarios. Across the first 32 scenarios the majority of the 1000 simulated datasets included only a handful of studies reporting only zero events. Therefore, in the final scenario 33 we further decreased the control group risk to lie between 0.1% and 0.3% in order to generate more studies with zero events in all treatment groups. Overall, though, the impact of such studies on the final results cannot be fully captured through the present

TABLE 1 Overview of scenarios examined in our simulation. For each scenario we generated 1000 datasets. The rows in bold represent the scenarios with multi-arm studies

#	Treatments in the network	Patients per arm	Number of studies per comparison	Total number of studies per dataset	Heterogeneity (τ)	Control group risk (%)	Mean events per study
1	5	30–60	2	20	0	3%–5%	3
2	5	30–60	2	20	0.1	3%–5%	3
3	5	30–60	2	20	0	5%–10%	6
4	5	30–60	2	20	0.1	5%–10%	6
5	5	30–60	4	40	0	3%–5%	3
6	5	30–60	4	40	0.1	3%–5%	3
7	5	30–60	4	40	0	5%–10%	6
8	5	30–60	4	40	0.1	5%–10%	6
9	8	30–60	2	56	0	3%–5%	3
10	8	30–60	2	56	0.1	3%–5%	3
11	5	100-200	2	20	0	1%–2%	4
12	5	100–200	2	20	0.1	1%–2%	4
13	5	100–200	2	20	0	0.5%–1%	2
14	5	100–200	2	20	0.1	0.5%–1%	2
15	5	100–200	4	40	0	0.5%–1%	2
16	5	100–200	4	40	0.1	0.5%–1%	2
17	**3**	**100–200**	**8**	**8**	**0**	**1%–2%**	**4**
18	**3**	**100–200**	**8**	**8**	**0.1**	**1%–2%**	**4**
19	**3**	**100–200**	**8**	**8**	**0**	**0.5%–1%**	**2**
20	**3**	**100–200**	**8**	**8**	**0.1**	**0.5%–1%**	**2**
21	**3**	**100–200**	**8**	**8**	**0**	**0.5%–5%**	**7**
22	**3**	**100–200**	**8**	**8**	**0.1**	**0.5%–5%**	**7**
23	**3**	**100–200**	**8**	**8**	**0**	**0.5%–10%**	**13**
24	**3**	**100–200**	**8**	**8**	**0.1**	**0.5%–10%**	**13**
25	**5**	**100–200**	**8**	**8**	**0**	**1%–2%**	**4**
26	**5**	**100–200**	**8**	**8**	**0.1**	**1%–2%**	**4**
27	**5**	**100–200**	**8**	**8**	**0**	**0.5%–1%**	**2**
28	**5**	**100–200**	**8**	**8**	**0.1**	**0.5%–1%**	**2**
29	**5**	**100–200**	**8**	**8**	**0**	**0.5–5%**	**7**
30	**5**	**100–200**	**8**	**8**	**0.1**	**0.5%–5%**	**7**
31	**5**	**100–200**	**8**	**8**	**0**	**0.5%–10%**	**13**
32	**5**	**100–200**	**8**	**8**	**0.1**	**0.5%–10%**	**13**
33	5	100–200	4	40	0	0.1%–0.3%	1

simulation study and further work is necessary. The extent of all-zero event studies can be found in Table 3 of our Supplementary material S1.

3.2 | Evaluated models

The simulation was conducted using R version 4.0.3 (October 10, 2020) and the packages netmeta,[27] brglm,[28] lme4,[29] and gemtc.[30] The packages netmeta[27] and gemtc[30] are tailored for frequentist and Bayesian NMA, respectively. The packages lme4[29] and brglm[28] allow the conduction of NMA either as a generalized linear mixed effects model or as a generalized linear model, respectively. The code used for our simulation can be found at https://github.com/TEvrenoglou/PL-NMA. In each scenario, we generated 1000 datasets and we compared the following models:

- Common- and random-effects IV-NMA model with 0.5 correction for studies with at least one zero event arm
- Common-effect MH-NMA and NCH-NMA without 0.5 correction for studies with zero event arms. Note that the NCH-NMA model also allows for random effects. However, the version of this model used here is the one implemented in netmeta[27] which is only common-effect and uses Breslow's approximation.
- Common- and random-effects PL-NMA model.
- Common-effect logistic regression NMA model with standard (non-penalized) likelihood.
- Random-effects logistic regression NMA model using MLE with fixed study intercept (BN-NMA). The model assumes the exact binomial likelihood for the observed events within studies and a normal distribution for the random-effects across the studies. The glmer function from the lme4[29] package was used to perform NMA. This function is not tailored for NMA and does not meet the requirement that variance-covariance matrix of the random-effects is having the structure with τ^2 for diagonal elements and $\frac{\tau^2}{2}$ for off-diagonal elements. The latter implies that this function can only be used for scenarios 1–16 where NMA of only two arm studies takes place.
- Common and random-effects Bayesian NMA with exact binomial likelihood and non-informative priors for the treatment effects. The common-effect model assumes that $d_{b_i k} \sim N\left(0, 100^2\right)$. For random-effects the study-specific treatment effects are following a normal distribution $\delta_{i,b_i k} \sim N\left(d_{b_i k}, \tau^2\right)$, then we form two random-effects models: the first assumes that $d_{b_i k} \sim N\left(0{,}100^2\right)$ and for the heterogeneity parameter that $\tau \sim \mathit{Unif}(0, 2)$.[31] The second random-effects model assumes narrower priors; $d_{b_i k} \sim N\left(0, 10^2\right)$ and a half-normal distribution for $\tau \sim \mathit{HN}(1)$. For all Bayesian models, we run 2 chains with 50 000 iterations and discarded the first 10 000 samples from each chain. Convergence was checked using the Brooks–Gelman–Rubin criterion[32] with a value lower than 1.1 considered as indicating convergence.

The objective of scenario 33 was to investigate the impact of all-zero event studies. Therefore, in this scenario we excluded the IV-NMA, the MH-NMA, and the NCH-NMA models as those models need to exclude all-zero event studies prior to the analysis.[4] The rest of the models were evaluated once by including and once by excluding all-zero event studies. A brief description of all models that are not described in the article is provided in the Supplementary material S1.

3.3 | Estimands and performance measures

The estimands of interest are the $T-1$ *logORs* between treatments $t = 2, \ldots, T$ and treatment 1. The performance of the models was investigated in terms of the mean bias defined as the mean difference between the estimated and the true *logORs* averaged over simulated datasets and estimands for each NMA method: the true estimands are all positive so averaging over estimands makes sense. For heterogeneity parameter in terms of the models assuming additive heterogeneity terms we calculated the mean difference between the estimated and the true τ averaged over simulated datasets while for the multiplicative PL-NMA model we report the percentage of the simulated datasets for which $\widehat{\varphi} > 1$. We further calculated the mean coverage in each scenario defined as the percent of the corresponding 95% confidence or credible intervals that included the corresponding true *logOR*. Finally, the methods were compared in terms of the mean squared error (MSE) and length of confidence or credible intervals averaged across multiple contrasts and simulated datasets.

3.4 | Simulation results

In the following sections we present the results of our simulation across the different performance measures. We first discuss the results for scenarios 1–32 and then we discuss separately the results of scenario 33. Overall, the Bayesian random-effects model with the uniform prior for the heterogeneity parameter had a rather poor performance and therefore its results are not discussed in the text but only presented in the figures and the tables; the results of all other models are discussed in the following sections.

3.4.1 | Bias in the estimation of the *logORs*

In Figure 1, we present the results in terms of mean bias across multiple contrasts (multiple estimands). In most scenarios, IV-common-effect and IV-random-effects gave the most biased results. The two PL-NMA models overall produced the least biased results across scenarios. The MH-NMA, NCH-NMA, and BN-NMA models also performed generally well, but the MH-NMA and NCH-NMA resulted in large bias in scenarios with many treatments in the network and a small number of studies per comparison (ie, scenarios 13, 14). The performance of NCH model for the scenarios with higher risks (scenarios 3, 4, 7, 8, 23, 24, 31, 32) was decreased as this model uses Breslow's approximation which is valid only for very rare events (see the Supplementary material S1 for details). Interestingly, the PL-NMA models had a stable performance across the different scenarios in terms of bias with a maximum value of mean bias equal to 0.02.

The BN-NMA model provided satisfactory results in terms of bias in almost all scenarios with a maximum mean bias of 0.06. The performance of BN-NMA was increased in scenarios involving relatively larger studies, namely, when the number of participants per arm was between 100 and 200. The results of the common-effect logistic regression model with the standard (non-penalized) likelihood were equivalent to the results of the BN-NMA model. The random-effects Bayesian model with half-normal prior distribution for heterogeneity had an improved performance in terms of bias and a better performance in comparison to the common-effect Bayesian model with bias ranging from −0.03 to 0.04. In all scenarios, the results in terms of bias were not affected materially for any of the models when the data-mechanism involved heterogeneity. Overall, the IV-NMA, the MH-NMA, and the NCH-NMA models had the tendency to give negative bias whether the rest of the models were positively biased. Full results are available in the Supplementary Tables 1 and 2 and are consistent with the results that we obtained for each contrast separately (Supplementary Figures 1–4).

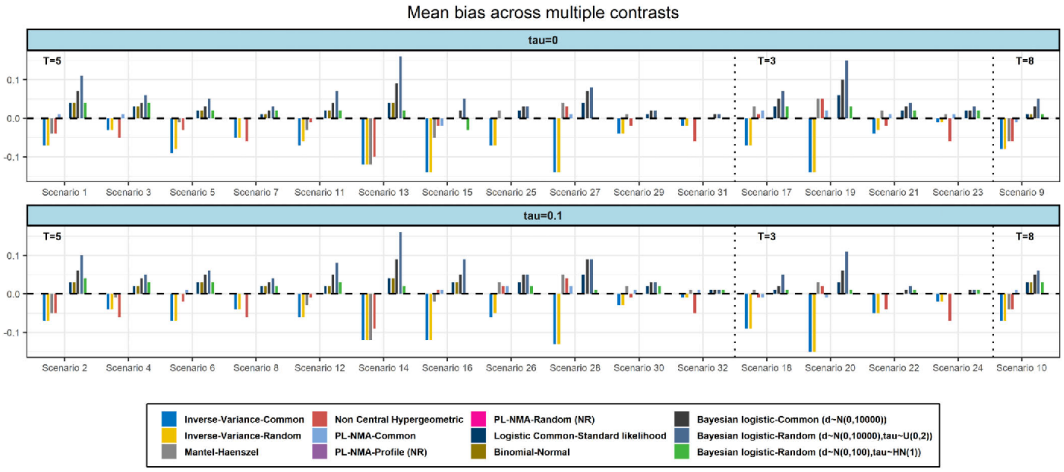


FIGURE 1 Simulation results in terms of mean bias for scenarios with 5, 3, and 8 treatments (T), respectively. Missing bars correspond to 0 mean bias (after rounding to two decimal places) for the respective NMA method. The Monte-Carlo standard error across the different scenarios and methods ranges from 0.004 to 0.02 with a mean value equal to 0.01. Models marked as *NR* are not relevant to the figure and thus no results are plotted

3.4.2 | Estimation of heterogeneity parameter

In terms of estimation of the heterogeneity, all the random-effects models appeared to have a bad performance (Table 2). The IV random-effects model seemed to have a particularly poor performance for estimating $\hat{\tau} \neq 0$. Especially for scenario 16, where 3 treatments and 8 studies per comparison were available, the model estimated that $\hat{\tau} = 0$ across 996 out of 1000 simulated datasets while the true value was set equal to 0.1. The BN-NMA model had also a very poor performance in estimating τ as in all heterogeneous scenarios the heterogeneity parameter was estimated being 0. Finally, since there is no clear correspondence between τ and φ we calculated the percentage of times with $\hat{\varphi} > 1$ across all datasets and scenarios. Across scenarios assuming heterogeneity ($\tau = 0.1$), $\hat{\varphi}$ was greater than 1 in a range between 0%–40% across the simulated datasets and scenarios; this percentage was increased to 20%–40% in scenarios with higher control group risks (scenarios 4, 8, 22, 24, 30, 32). Across scenarios assuming that $\tau = 0$ the percentage that $\hat{\varphi} > 1$ ranges from 0-36%. The percentage per scenario can be found in Table 2. The half-normal Bayesian random-effects model appeared to provide the most biased estimates of the heterogeneity parameter with bias ranging from 0.07 to 0.25.

3.4.3 | Coverage probability

In terms of coverage probability, the PL-NMA model with respect to Wald-type confidence interval had a consistent performance with coverage ranging from 93.3% to 96.3% in 18 scenarios the coverage was slightly below the nominal level but almost always within 95% confidence interval constructed for the nominal level (Figure 2). This confidence interval was calculated as $0.95 \pm 1.96\sqrt{\frac{0.95(1-0.95)}{1000}} = (0.936, 0.964) \times 100\%$. The performance of the PL-NMA model was increased by using profile likelihood confidence intervals and, in most scenarios, it was around the nominal level of 95% with a range between 94.3% and 96%. MH-NMA and NCH-NMA models had similarly good performance with coverage probability close to the nominal level, but it should be noted that the performance of NCH-NMA model was again decreased for some scenarios with higher control group risks (scenarios 23, 24, 31). These results are consistent with previous simulation studies.[4] The IV-NMA models had the worst performance as those models were suffering from over-coverage across almost all scenarios. Overall, for IV-NMA, MH-NMA, and NCH-NMA results in terms of coverage are consistent with previous simulation studies.[4] The BN-NMA model and two Bayesian models had in most scenarios a good performance with a coverage probability around the nominal level and typically within confidence interval bounds for the nominal level.

3.4.4 | MSE and length of confidence or credible intervals

In terms of MSE all the models appeared to have a similar performance (Figure 3). The two Bayesian models had a slightly increased MSE which overall found to range from 0.07 to 0.72. Regarding the mean length of confidence (or credible intervals) the Wald-type confidence intervals obtained by the PL-NMA were found to have the smaller lengths in comparison to the other models (Figure 4). The mean length of the profile-likelihood confidence intervals for PL-NMA, though, were wider; this probably explains also their slightly better performance compared to the Wald-type intervals in terms of coverage probability. As expected, the most uncertain results were obtained by the Bayesian random-effects model which provided the widest credible intervals.

3.4.5 | Impact of all-zero event studies

In 6 datasets, all other models but the PL-NMA models provided implausible results with standard errors larger than 10^8. Thus, all performance measures were calculated for all models after excluding those 6 scenarios.

For the PL-NMA model, the inclusion or exclusion of all-zero event studies do not seem to have important impact to the point estimates of the *logORs*, thus the results do not differ much in terms of bias. Bayesian models seem to have a poor performance, irrespective of the choice between inclusion or exclusion. In particular, we found that including all-zero events in the Bayesian models can seriously bias the results. The decision between

TABLE 2 Results in terms of bias for the estimation of heterogeneity across all scenarios. For the case of PL-NMA model the frequency that $\hat{\varphi} > 1$ is reported instead of the bias. In parenthesis the percentage that $\hat{\tau} > 0$ or $\hat{\varphi} > 1$ across the 1000 datasets and 32 scenarios. In bold the scenarios that assume $\tau = 0.1$

#	IV-random Mean bias (% $\hat{\tau}$ >0)	Binomial normal Mean bias (% $\hat{\tau}$ >0)	Bayesian $d \sim N(0, 100^2)$, $\tau \sim U(0, 2)$ Mean bias (% $\hat{\tau}$ >0)	Bayesian $d \sim N(0, 10^2)$, $\tau \sim HN(1)$ Mean bias (% $\hat{\tau}$ >0)	PL-NMA-random Frequency $\hat{\varphi} > 1$ (% $\hat{\varphi} > 1$)
1	0.05 (15.5%)	0 (0%)	0.54 (100%)	0.25 (100%)	94 (9.4%)
2	**−0.04 (16.5%)**	**−0.1 (0%)**	**0.46 (100%)**	**0.15 (100%)**	**101 (10.1%)**
3	0.09 (29.6%)	0 (0%)	0.36 (100%)	0.22 (100%)	246 (24.6%)
4	**0.01 (37%)**	**−0.1 (0%)**	**0.28 (100%)**	**0.13 (100%)**	**290 (29%)**
5	0.01 (5.5%)	0 (0%)	0.37 (100%)	0.23 (100%)	21 (2.1%)
6	**−0.08 (7.5%)**	**−0.1 (0%)**	**0.29 (100%)**	**0.13 (100%)**	**33 (3.3%)**
7	0.05 (22.8%)	0 (0%)	0.26 (100%)	0.20 (100%)	155 (15.5%)
8	**−0.03 (29.9%)**	**−0.1 (0%)**	**0.18 (100%)**	**0.11 (100%)**	**202 (20.2%)**
9	0.01 (3.3%)	0 (0%)	0.34 (100%)	0.23 (100%)	10 (1%)
10	**−0.09 (5.6%)**	**−0.1 (0%)**	**0.25 (100%)**	**0.13 (100%)**	**19 (1.9%)**
11	0.05 (17%)	0 (0%)	0.44 (100%)	0.23 (100%)	112 (11.2%)
12	**−0.03 (21.7%)**	**−0.1 (0%)**	**0.36 (100%)**	**0.14 (100%)**	**141 (14.1%)**
13	0.01 (2.6%)	0 (0%)	0.71 (100%)	0.25 (100%)	7 (0.7%)
14	**−0.09 (3.9%)**	**−0.1 (0%)**	**0.62 (100%)**	**0.16 (100%)**	**12 (1.2%)**
15	0.00 (0.3%)	0 (0%)	0.50 (100%)	0.24 (100%)	0 (0%)
16	**−0.10 (0.4%)**	**−0.1 (0%)**	**0.41 (100%)**	**0.14 (100%)**	**0 (0%)**
17	0.06 (18.4%)	-	0.43 (100%)	0.23 (100%)	127 (12.7%)
18	**−0.03 (24.2%)**	-	**0.34 (100%)**	**0.14 (100%)**	**161 (16.1%)**
19	0.01 (4.8%)	-	0.65 (100%)	0.25 (100%)	18 (1.8%)
20	**−0.08 (5.1%)**	-	**0.55 (100%)**	**0.15 (100%)**	**30 (3%)**
21	0.07 (27.8%)	-	0.32 (100%)	0.21 (100%)	201 (20.1%)
22	**−0.01 (33.1%)**	-	**0.23 (100%)**	**0.11 (100%)**	**236 (23.6%)**
23	0.08 (38.2%)	-	0.24 (100%)	0.18 (100%)	321 (32.1%)
24	**−0.01 (41.9%)**	-	**0.15 (100%)**	**0.09 (100%)**	**356 (35.6%)**
25	0.03 (14.8%)	-	0.31 (100%)	0.21 (100%)	92 (9.2%)
26	**−0.06 (18.1%)**	-	**0.22 (100%)**	**0.12 (100%)**	**120 (12%)**
27	0.00 (1.7%)	-	0.47 (100%)	0.24 (100%)	11 (1.1%)
28	**−0.09 (2.3%)**	-	**0.37 (100%)**	**0.14 (100%)**	**10 (1%)**
29	0.05 (25.3%)	-	0.23 (100%)	0.18 (100%)	167 (16.7%)
30	**−0.03 (32.9%)**	-	**0.15 (100%)**	**0.09 (100%)**	**238 (23.8%)**
31	0.06 (34.8%)	-	0.17 (100%)	0.15 (100%)	276 (27.6%)
32	**−0.01 (47.6%)**	-	**0.09 (100%)**	**0.07 (100%)**	**387 (38.7%)**

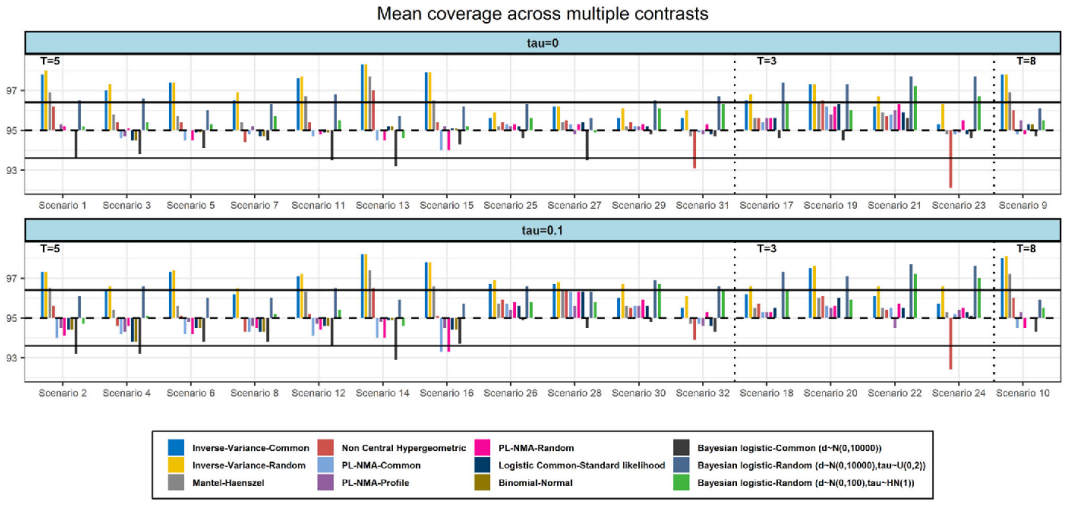


FIGURE 2 Simulation results in terms of coverage probability for scenarios with 5, 3, and 8 treatments (T), respectively. The horizontal lines represent the upper and lower bounds of the 95% confidence interval which is constructed for the nominal level

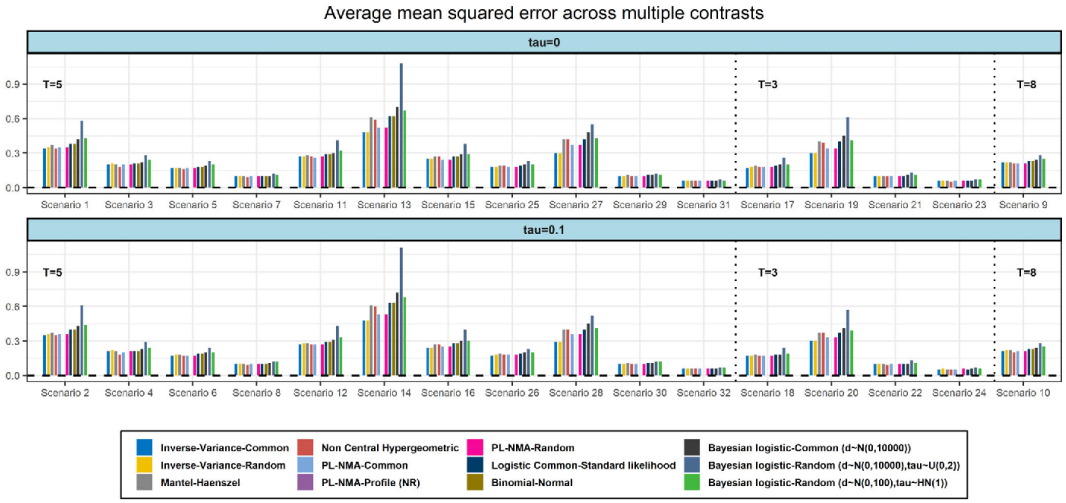


FIGURE 3 Simulation results in terms of MSE for scenarios with 5, 3, and 8 treatments (T), respectively. Models marked as *NR* are not relevant to the figure and thus no results are plotted

inclusion or exclusion of such studies appeared to have no impact for the BN-NMA and common-effect logistic regression model. Those models do not require such studies to be excluded prior to the analysis but they intrinsically exclude them.[8]

The PL-NMA model and BN-NMA estimated that $\hat{\varphi} = 1$ and $\hat{\tau} = 0$ across all the 1000 datasets of scenario 33 in both cases of inclusion or exclusion of all-zero event studies. The Bayesian random-effects model provided highly biased estimates of the heterogeneity parameter with mean bias equal to 0.26 and 0.27 when including and excluding all-zero event studies, respectively.

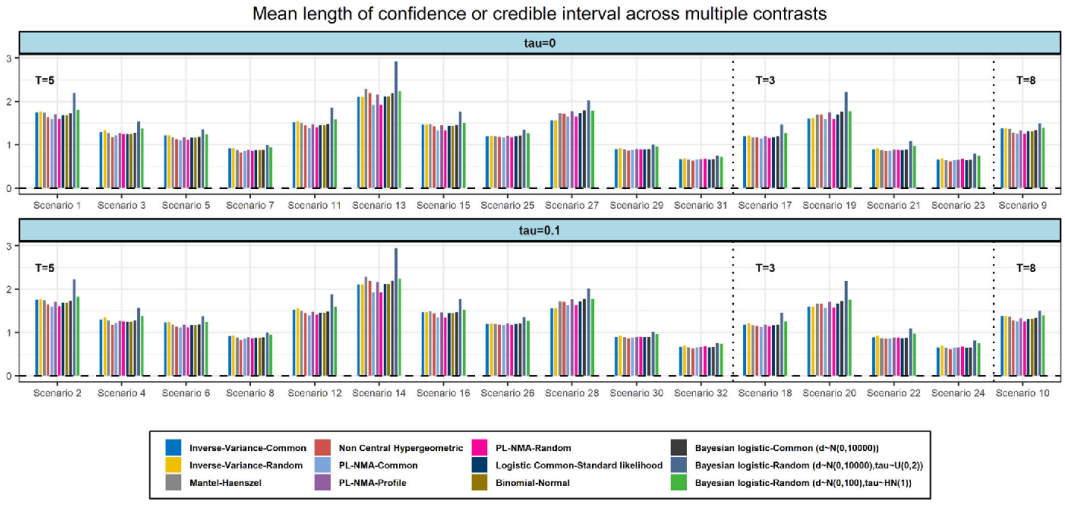


FIGURE 4 Simulation results in terms of length of confidence (or credible) intervals for scenarios with 5, 3, and 8 treatments (T), respectively

Inclusion of all-zero event studies seemed to slightly improve the performance of PL-NMA model in terms of coverage probability. However, Wald-type intervals tended to suffer from under-coverage irrespectively of including or excluding while profile-likelihood confidence intervals tended to provide coverage above the nominal level of 95%. Finally, the Bayesian models were suffering from under-coverage; though their performance was improved when all-zero event studies were included.

For the PL-NMA model the MSE was slightly decreased when including all-zero event studies while the Wald-type confidence intervals were generally smaller. The mean length of profile likelihood confidence intervals remained robust and were not affected by the inclusion or the exclusion. For the Bayesian models the results did not differ when including or excluding all-zero event studies. More details about the results of this scenario can be found in the Supplementary material S1.

4 | CLINICAL EXAMPLES

4.1 | Safety of inhaled medications for patients with chronic obstructive pulmonary disease

We compared the results across the different models using a network that evaluates the safety of inhaled medications for patients with chronic obstructive pulmonary disease.[33] The network consists of 41 studies and the outcome of interest is mortality. This is supposed to be a rare outcome as out of 52 462 patients only 2408 (4.6%) experienced the event of interest. In total, 13 out of 41 studies reported less than 5 events and 20 out of 99 arms had 0 events. The network diagram of this dataset is depicted in Figure 5. We analyzed the data using the PL-NMA model both with Wald and profile likelihood CIs, the MH-NMA model, the NCH-NMA model and finally the same Bayesian models that were included in our simulation For the latter convergence was checked using the Brooks–Gelman–Rubin[32] criterion. We did not use the IV model and the Bayesian random-effects model with uniform heterogeneity prior due to the bad performance of these models in the simulation.

The results across the different NMA methods are almost identical (Figure 6). Regarding the estimation of heterogeneity, the additive parameter τ was estimated as zero for the BN-NMA model and the multiplicative parameter φ was estimated as 1. However, the Bayesian random-effects model resulted in $\hat{\tau}=0.12$ [0, 0.35] for the model that assumes $HN(1)$ prior.

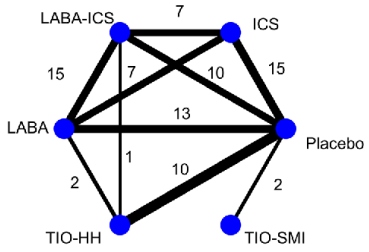


FIGURE 5 Network diagram for the inhaled medications example. HH, tiotropium dry powder; ICS, inhaled corticosteroid; LABA, long-acting β2 agonist; TIO-TIO-SMI, tiotropium solution

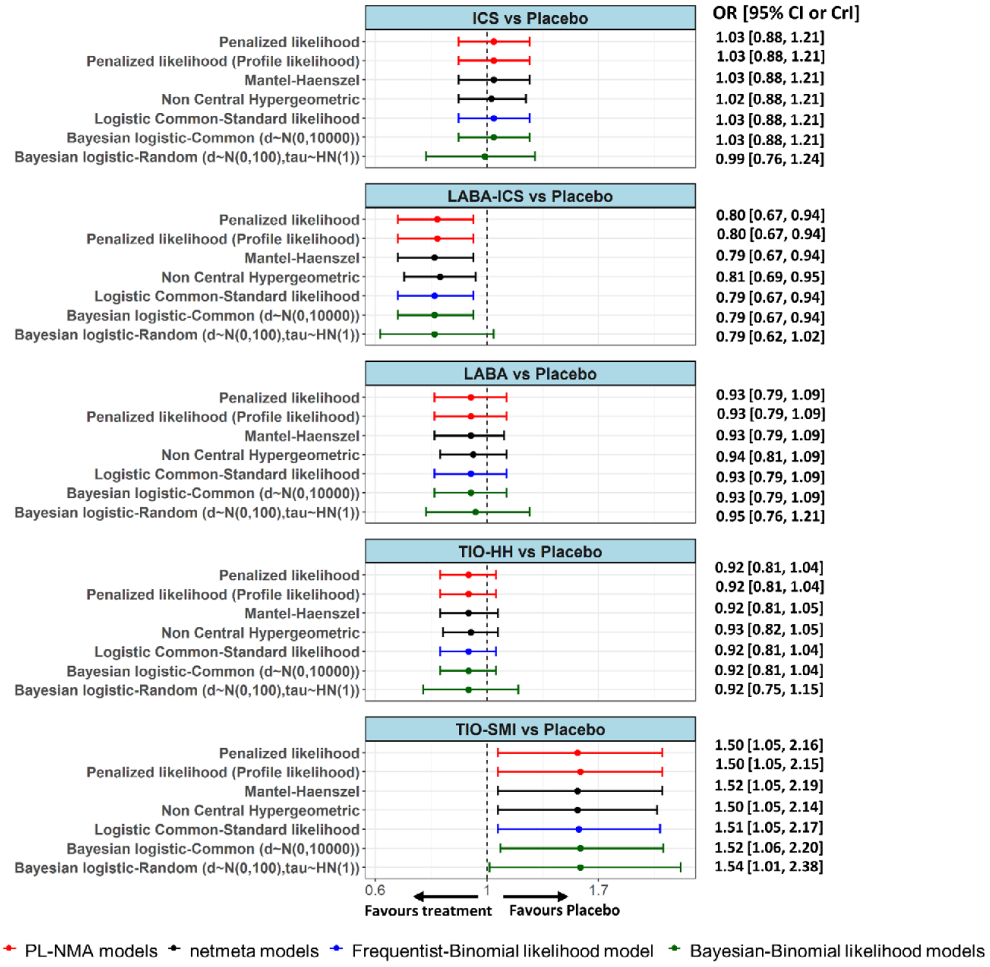


FIGURE 6 Forest plots showing the odds ratios obtained from the inhaled medications network for all comparisons against the reference

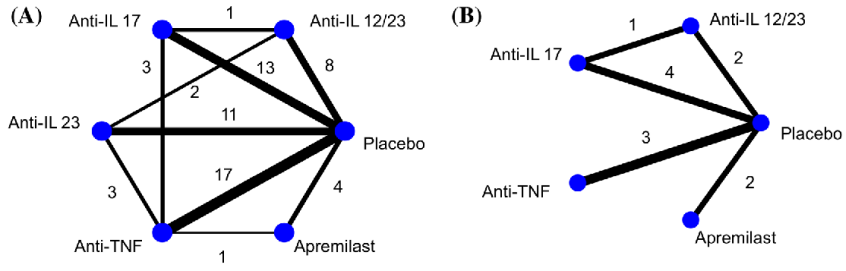


FIGURE 7 Network diagrams for the psoriasis example. Panel (A) shows the initially well-connected network while panel (B) the resulting network after the exclusion of studies that report only zero events. The node Anti-IL 23 is eliminated with the discarded studies. Anti-IL, anti-interleukin; Anti-TNF, anti-tumor necrosis factor

4.2 | Safety of different drug classes for chronic plaque psoriasis

The second example is a network that evaluates the safety of different interventions for chronic plaque psoriasis.[34] The dataset consists of 43 studies that involve 5 drug classes and placebo. Here, we compared again the different NMA methods but in a more extreme situation where the control group risks in the studies range between 0 and 1%. The outcome of interest is the number of malignancies that occurred after using the drugs. The mean sample size per study arm is 226 patients. Out of the 43 studies in this network, 15 (35%) are studies with zero events in all treatment groups. For MH-NMA and NCH-NMA the exclusion of all-zero event arms leads to the removal of the treatment node Anti-IL 23 from the network (Figure 7A,B). A more detailed explanation regarding the exclusion mechanism of those models can be found in our Supplementary material S1. The resulting network appears to be connected in terms of the rest treatments but also much reduced in total number of studies than the initial network.

The results of the different approaches are shown in Figure 8. In terms of point estimates, for comparison Anti-IL 12/23 vs Placebo all models except the Bayesian random-effects models gave very similar results. For the comparison Apremilast vs Placebo, the results across all models appeared to be robust. For the rest of the comparisons there is a lot of diversity across the estimated point estimates which depicts the sensitivity of the results when the events are very rare. Finally, in all comparisons there are important differences across the Bayesian models which validates our hypothesis that the results are strongly dependent on the choice of prior distribution. With respect to heterogeneity, the PL-NMA estimated the multiplicative term φ as 1. So, it suggests the absence of statistical heterogeneity. The Bayesian random-effects model, though, gave $\widehat{\tau}$= 0.56 (0.03, 1.61).

5 | DISCUSSION

In this article, we have presented a new method for NMA of binary outcomes and rare events. Our method aims to improve the performance of existing NMA approaches for rare events in terms of bias and precision by using the penalized likelihood function for logistic regression that was originally proposed by Firth[13] for individual studies. Our approach can only provide odds ratios. However, in the context of rare events, the differences between odds ratios and risk ratios are often negligible.[35,36]

We evaluated the performance of the different methods and compared their results through an extended simulation study and two real clinical examples. We used various scenarios including studies with low or extremely low control group risks. The proposed PL-NMA model appeared to have overall the best performance in terms of bias especially in situations with very few studies per comparison and very low control group risks. The performance on coverage probability was improved when profile-likelihood confidence intervals were used but precision was reduced. In agreement with previous simulation studies,[4] the IV-NMA models are considered a suboptimal choice and we believe meta-analysts should avoid their use for NMAs with rare events, especially for the cases with very small number of events per study (eg, <3). MH-NMA, NCH-NMA are good options under certain conditions, but they become less reliable as the network becomes sparser. BN-NMA model is shown to be robust in terms of bias. However, we only evaluated the performance

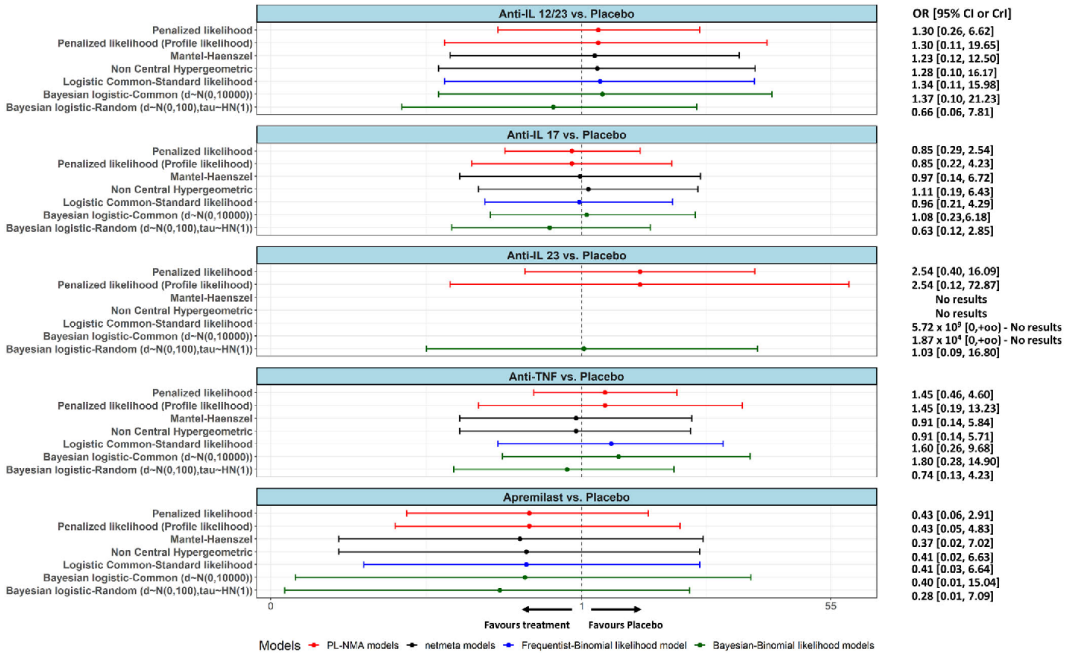


FIGURE 8 Forest plots showing the odds ratios obtained from psoriasis network for all comparisons against the reference

of the BN-NMA model in scenarios with two-arm studies as currently the model cannot take into account multivariate distributions for the random-effects. The common-effect Bayesian model performed well in terms of bias and coverage. The performance of the random-effects Bayesian model was much improved across all the performance measures when we moved to narrower prior distributions for both the treatment effects and the heterogeneity parameters. In that case the model provided a reliable alternative as the results were less biased in comparison to the common-effect Bayesian model and the coverage probability was usually above the nominal level. The estimates obtained from the penalized likelihood function are typically shrunk towards 0 and this in turn results in smaller standard errors. Hence, sometimes the PL-NMA may underestimate the true study variance. As expected, all frequentist models failed to sufficiently frequently detect the presence of heterogeneity particularly for scenarios with extremely low control group risks (ie, 0.5%–1% and 1%–2%). On the other hand, Bayesian models identified some amount of heterogeneity but the estimates were in all cases highly biased. Of course the latter is a consequence coming from the choice of the prior distribution and other more suitable choices can potentially give less biased results.

An additional property of the proposed PL-NMA model is the ability to synthesize all studies within a network of interventions irrespective of the number of events per arm since excluding such studies from the analysis may result in disconnected networks. The problem lies mainly in networks including studies with zero events in two or more treatment groups. To date, the optimal way to treat such studies in meta-analysis or NMA is still unclear with some researchers arguing that they are informative[11,37] and others that those studies are non-informative and problematic.[4,36] In scenario 33 of our simulation study, we intended to investigate the performance of the different models in the presence of several such studies. The choice of inclusion or exclusion appears to mostly affect the uncertainty measures such as the length of the confidence or credible intervals and the MSE which were typically smaller for all models when including all-zero event studies. In terms of bias, the inclusion or exclusion appeared to have no serious impact for the PL-NMA model. On the contrary, inclusion biased substantially the Bayesian models.

Although the PL-NMA model is by nature a common-effect model, we used a two-stage approach for incorporating between-study variance as a multiplicative overdispersion parameter. Multiplicative parameters are not the standard way to account for heterogeneity in meta-analysis as typically an additive parameter is implemented. However, the

penalized likelihood approach strongly relies on the analytical expression of the likelihood function and its moments; these are not available for the logistic regression model of Section 2.1 for the case of random-effects. Thus, the incorporation of an additive heterogeneity parameter in the PL-NMA method is not straightforward. The use of multiplicative heterogeneity parameters has been suggested previously for some meta-analytical approaches[22,23] and particularly for meta-analysis of rare events by Kuss[37] and by Kulinskaya and Olkin.[38] The latter, in contrast to our approach, is a one-stage random-effects model.[38] A key difference between our two-stage approach and the usual one-stage random-effects models is that the former does not affect the relative effect estimates but only inflates their variance when $\varphi > 1$. In addition, the incorporation of the multiplicative parameter does not affect the weights of the studies and, thus, results might be dominated by the larger studies.[24,39] On the other hand, the additive heterogeneity model downweights larger studies. Hence, the additive model may be more plausible when larger studies have more heterogeneity while the multiplicative when the smaller studies do so.[40,41] A potential interpretation of the multiplicative model could be that the scale parameter represents a common variance and the study-specific variances represent differing sample sizes. In our simulation, we found that all random-effects approaches did not perform sufficiently well with respect to the estimation of the heterogeneity. It should be noted, though, that the Cochrane handbook suggests that estimation of heterogeneity is not of primary interest when performing meta-analyzes of rare events and the priority should be the estimation of the treatment effects.[36]

Overall, the proposed PL-NMA model offers a reliable choice for performing NMA for binary outcomes with rare events. Future work for NMA of rare events could consider the extension of the beta-binomial model for meta-analysis into NMA.[37] Meta-analysts should always bear in mind that the presence of studies with rare events makes estimation challenging and, therefore, a sensitivity analysis should be carried out to investigate the robustness of the results under various analysis schemes. It is planned to integrate the proposed PL-NMA method into an R package.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Gerta Rücker for providing useful comments on an earlier version of the manuscript. Ian White was supported by the Medical Research Council Program MC_UU_00004/06. Dimitris Mavridis is funded by the European Union's Horizon 2020 COMPAR-EU project (No 754936).

DATA AVAILABILITY STATEMENT

Data and R codes for the analysis of the two clinical examples can be found in the Supplementary material.

ORCID

Theodoros Evrenoglou https://orcid.org/0000-0003-3336-8058
Ian R. White https://orcid.org/0000-0002-6718-7661
Sivem Afach https://orcid.org/0000-0003-2791-223X
Dimitris Mavridis https://orcid.org/0000-0003-1041-4592
Anna Chaimani https://orcid.org/0000-0003-2828-6832

REFERENCES

1. Higgins JP, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? *Lancet*. 2015;386(9994):628-630.
2. Efthimiou O, Debray TP, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. *Res Synth Methods*. 2016;7(3):236-263.
3. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29(29):3046-3067.
4. Efthimiou O, Rücker G, Schwarzer G, Higgins JPT, Egger M, Salanti G. Network meta-analysis of rare events using the mantel-Haenszel method. *Stat Med*. 2019;38(16):2992-3012.
5. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol*. 2009;9(1):56.
6. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ*. 2016;352:i1981.
7. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004;23(9):1351-1375.
8. Bradburn MJ, Deeks JJ, Berlin JA, Russell LA. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26(1):53-77.
9. Senn S. Hans van Houwelingen and the art of summing up. *Biom J*. 2010;52(1):85-94.
10. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059-1085.

11. Xu C, Li L, Lin L, et al. Exclusion of studies with no events in both arms in meta-analysis impacted the conclusions. *J Clin Epidemiol*. 2020;123:91-99.
12. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res*. 2016;25(6):2858-2877.
13. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
14. King G, Zeng L. Logistic regression in rare events data. *Polit Anal*. 2001;9(2):137-163.
15. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474. doi:10.1177/0962280218784726
16. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008;17(3):279-301.
17. Kosmidis I, Firth D. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*. 2021;108(1):71-82.
18. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409-2419.
19. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1-10.
20. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. *Am J Epidemiol*. 2018;187(4):864-870.
21. White IR, Turner RM, Karahalios A, Salanti G. A comparison of arm-based and contrast-based models for network meta-analysis. *Stat Med*. 2019;38:5197-5213. doi:10.1002/sim.8360
22. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. *Stat Med*. 2015;34(13):2116-2127.
23. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-regression. *Res Synth Methods*. 2017;8(1):19-42.
24. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med*. 1999;18(20):2693-2708.
25. Fletcher DJ. Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika*. 2012;99(1):230-237.
26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
27. Rücker G, Krahn U, König J, Efthimiou O, Schwarzer G. Netmeta: network meta-analysis using frequentist methods; 2020. https://CRAN.R-project.org/package=netmeta
28. Kosmidis I. brglm: Bias reduction in binary-response generalized linear models. R package version 071; 2020. https://cran.r-project.org/package=brglm
29. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;1(1):1-48.
30. van Valkenhoef G, Lu G, de Brock B, Hillege H, Ades AE, Welton NJ. Automating network meta-analysis. *Res Synth Methods*. 2012;3(4):285-299.
31. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network Meta-Analysis for Decision-Making*. Chichester, UK: John Wiley & Sons; 2018.
32. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*. 1998;7(4):434-455.
33. Dong Y-H, Lin H-H, Shau W-Y, Wu Y-C, Chang C-H, Lai M-S. Comparative safety of inhaled medications in patients with chronic obstructive pulmonary disease: systematic review and mixed treatment comparison meta-analysis of randomised controlled trials. *Thorax*. 2013;68(1):48-56.
34. Afach S, Chaimani A, Evrenoglou T, et al. Meta-analysis results do not reflect the real safety of biologics in psoriasis. *Br J Dermatol*. 2021;184(3):415-424.
35. Efthimiou O. Practical guide to the meta-analysis of rare events. *Evid Based Ment Health*. 2018;21(2):72-76.
36. Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: Wiley; 2019:241-284.
37. Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Stat Med*. 2015;34(7):1097-1116.
38. Kulinskaya E, Olkin I. An overdispersion model in meta-analysis. *Stat Model*. 2014;14(1):49-76.
39. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet*. 1991;338(8775):1127-1130.
40. Schmid CH. Heterogeneity: multiplicative, additive or both? *Res Synth Methods*. 2017;8(1):119-120.
41. Mawdsley D, Higgins JP, Sutton AJ, Abrams KR. Accounting for heterogeneity in meta-analysis using a multiplicative model-an empirical study. *Res Synth Methods*. 2017;8(1):43-52.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Evrenoglou T, White IR, Afach S, Mavridis D, Chaimani A. Network meta-analysis of rare events using penalized likelihood regression. *Statistics in Medicine*. 2022;1-17. doi: 10.1002/sim.9562

3. Part 2: Dealing with rare events in meta-analysis of Covid-19 trials

In March 2020 the outbreak of the coronavirus disease 2019 (Covid-19) was declared as a pandemic by the WHO. This event motivated the researchers all over the world to respond to this emergency situation and thus to rapidly develop and evaluate preventive and therapeutic interventions for treating Covid-19. Given the unprecedented explosion of clinical findings and the inevitable uncertainty which accompanies them several living evidence synthesis initiatives[106–110] were set up to provide up to date evidence about the effect of the different interventions for Covid-19.

The COVID-NMA initiative[106–108] was set up to perform a living systematic review of Covid-19 trials with three main tasks. At first, to continuously collect and to critically appraise all the available evidence related to Covid-19 interventions. At second, to synthesize the respective findings through meta-analysis and finally at third to make the results freely and publicly available through the COVID-NMA platform (https://covid-nma.com/). Up to September 2022 the COVID-NMA database for pharmacologic interventions was consisted by 533 RCTs which are investigating in total more than 300 pairwise comparisons. Moreover, the corresponding database for Covid-19 vaccines consisted of 32 RCTs investigating the efficacy and safety of vaccines which are related to 23 developers.

According to the COVID-NMA's protocol the process of the evidence synthesis takes place through the frequentist IV random-effects model with the heterogeneity parameter being estimated by using the REML method[107,108]. However, with such an amount of available data it is inevitable

that some special meta-analysis topics like rare events will appear and thus potentially threaten the validity of the findings[41]. Rare events frequently appear across many of the primary and safety outcomes which are investigated by COVID-NMA. For example, for the outcome of serious adverse events (SAE) the corresponding risk of event across studies involving hospitalized patients is limited to only 8% and it can further reduce to 4% if studies with outpatients are considered as well. Moreover, around 25% of RCTs investigating SAE report zero events in all their intervention arms. As it was explained in chapters 1 and 2 in such cases of sparsity the IV model is expected to give biased summary results while the conduction of meta-analysis can even be pointless if many of the studies report zero events in one or all their intervention arms. The estimation of the heterogeneity parameter based on the REML method can be problematic as the resulting estimates are biased in presence of rare events with τ^2 often being estimated as 0[34,35]. Moreover, the REML method can even fail to provide estimates for heterogeneity as it relies on an iterative process which can face convergence issues in cases of sparse data.

Beyond the special analysis topic of rare events some additional issues can arise regarding the posting of the results through solely a platform. For instance, guideline developers and other stakeholders, apart from real-time access to high-quality data, also need to be able to investigate the data and the impact of different characteristics on the results as well as to produce their preferred evidence summaries. Accommodating such preferences of different end-users of the platform is not feasible in the form of pre-conducted non amendable meta-analyses, given the high number of comparisons included in the living review. In particular, producing and publishing online all possible sub-analyses for each comparison a) is extremely time-consuming and b) would make the platform cumbersome for the users.

To address all the aforementioned important points, the metaCOVID application was developed and made freely available through: https://covid-nma.com/metacovid/. This is an R-Shiny web-application which allows all the end-users of the COVID-NMA platform and other external researchers to directly perform their own meta-analysis of Covid-19 trials based on the most up to date data. The basic advantage of the metaCOVID is that through its user-friendly environment the users can investigate complex data structures and fit several meta-analysis models without any requirement of programming expertise. This application provides the flexibility of replacing the IV model with other more suitable models for the analysis of rare events. Specifically, the IV model can be replaced with either the non-parametric methods of MH and Peto or with the penalized likelihood model which was described in the chapter 2 of this thesis. Even though the penalized likelihood model was initially proposed as a NMA model it can be easily reduced with only minor modifications to account for pairwise meta-analysis. The latter is important as currently metaCOVID follows the plan of the COVID-NMA initiative and thus it only performs pairwise meta-analyses. Finally, the penalized likelihood model can facilitate the calculation of summary intervention effects even in extreme cases where the most or all the RCTs report zero events in one or all of their arms.

Additional analysis options are available through metaCOVID and include a variety of options for the user in terms of the method which is used for estimating the heterogeneity parameter τ^2. Specifically, metaCOVID allows the replacement of the REML method[22] with one of the following methods: the DerSimonian-Laird estimate[23], the Paul-Mandel estimate[24], the Sidik-Jonkman estimate[40] and the estimator that is based on empirical Bayes inference[111]. Finally, metaCOVID can have a broader scope than simply analyzing rare events. The application allows for several sensitivity analyses to be conducted for example by excluding or not RCTs labelled with high risk

of bias or by excluding preprint publications. Additionally, several subgroup analyses can be conducted based on trial characteristics such as their funding or location status.

The detailed presentation about the metaCOVID application is given by the following manuscript: Evrenoglou, T., Boutron, I., Seitidis, G., Ghosn, L. and Chaimani, A. (2023), metaCOVID*:* A web-application for living meta-analyses of COVID-19 trials. Res Syn Meth. Accepted Author Manuscript. https://doi.org/10.1002/jrsm.1627

Software freely available online https://covid-nma.com/metacovid/

Evrenoglou Theodoros (Orcid ID: 0000-0003-3336-8058)
Seitidis Georgios (Orcid ID: 0000-0003-0856-1892)
Chaimani Anna (Orcid ID: 0000-0003-2828-6832)

metaCOVID: A web-application for living meta-analyses of COVID-19 trials

Theodoros Evrenoglou*[1], Isabelle Boutron[1,2], Georgios Seitidis[3], Lina Ghosn[1,2], Anna Chaimani[1]

[1] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Centre of Research in Epidemiology and Statistics (CRESS-U1153), F-75004 Paris, France

[2] Centre d'Épidémiologie Clinique, AP-HP, Hôpital Hôtel-Dieu, F-75004 Paris, France

[3] Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

*Corresponding Author:

Theodoros Evrenoglou

Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Centre of Research in Epidemiology and Statistics (CRESS-U1153), F-75004 Paris, France

Hôpital Hôtel-Dieu, 1 Place du Parvis Notre-Dame, 75004 Paris

email: tevrenoglou@gmail.com

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1627

Abstract

Abstract: Outputs from living evidence syntheses projects have been used widely during the pandemic by guideline developers to form evidence-based recommendations. However, the needs of different stakeholders cannot be accommodated by solely providing pre-defined non amendable numerical summaries. Stakeholders also need to understand the data and perform their own exploratory analyses. This requires resources, time, statistical expertise, software knowledge as well as relevant clinical expertise to avoid spurious conclusions. To assist them, we created the *metaCOVID* application which, based on automation processes, facilitates the fast exploration of the data and the conduct of sub-analyses tailored to end-users needs. *metaCOVID* has been created in R and is freely available as an R-Shiny application. Based on the COVID-NMA platform (https://covid-nma.com/) the application conducts living meta-analyses of RCTs related to COVID-19 treatments and vaccines for several outcomes. Several options are available for subgroup and sensitivity analyses. The results are presented in downloadable forest plots. We illustrate *meta COVID* through three examples involving well-known treatments and vaccines for COVID-19. The application is freely available from https://covid-nma.com/metacovid/.

Keywords: dynamic evidence synthesis; medical decision making; online software tool; data visualization

1 Introduction

The emergent situation of the COVID-19 pandemic motivated researchers worldwide to rapidly develop and evaluate preventive and therapeutic interventions. In light of an unprecedented explosion of clinical research findings and the uncertainty they are inevitably accompanied with, several initiatives (1–5) were set up to provide 'living' evidence on the effects of the different interventions; this means to continuously collect and synthesize all available evidence on COVID-19 treatments, preventive interventions, and vaccines.

The COVID-NMA platform (1–3) (https://covid-nma.com/) provides public access to the most up-to-date information with respect to the effects of the different COVID-19 interventions and supports timely clinical decisions and policy-making. However, guideline developers and other stakeholders, apart from real-time access to high-quality data, also need to be able to investigate the data and the impact of different characteristics on the results as well as to produce their preferred evidence summaries. For example, different countries have different policies concerning the inclusion or not of preprints (6) when forming COVID-19 guidelines while others prefer to exclude, by default, small trials (e.g. with less than 100 participants). Accommodating such preferences of different end-users of the platform is not feasible in the form of pre-conducted non-amendable meta-analyses, given the high number of comparisons included in the living review. In particular, producing and publishing online all possible sub-analyses for each comparison a) is extremely time-consuming and b) would make the platform cumbersome for the users.

On the other hand, it is of great importance to reassure that all sub-analyses undertaken are based on clinical rational. It is well-known that the larger the number of subgroup analyses conducted the higher the risk of obtaining false-positive results (7). A database of more than 400 variables, like the COVID-NMA database, is always subject to this phenomenon and should be used very carefully from investigators with understanding and knowledge of the data. Therefore, within COVID-NMA, only a small number of such secondary analyses are considered; these are

pre-defined exploratory analyses included in the COVID-NMA protocol chosen on the basis of the clinical expertise of the steering committee (1,2).

To address all these important points, we developed and made freely available the *'metaCOVID'* application (https://covid-nma.com/metacovid/). This web-application allows the end-users of the COVID-NMA platform and other external researchers to directly use the most up-to-date database and perform meta-analyses tailored to their needs in a user-friendly environment. The results of all analyses are summarized in the form of downloadable forest plots. A key feature of the application is that the numerical results are presented alongside study characteristics and risk of bias assessments. In this article, we describe the functionality of *metaCOVID* and the different available options and sub-analyses it offers. We also discuss how it can inform guideline developers about the impact of difference characteristics on the results using examples from the COVID-NMA data.

2 Methods

2.1 Data embedded in metaCOVID

The application is directly connected to the COVID-NMA database which is updated weekly or bi-weekly with data from all new RCTs evaluating COVID-19 treatments and vaccines. The search is conducted on a daily basis using the L.OVE (8) and the Cochrane COVID-19 Study register platforms. As of March 1, 2022, the COVID-NMA revised its protocol for treatments to update only comparisons evaluating immunomodulators and antiviral therapies. Every RCT in the database is also accompanied by domain-specific RoB assessments (for randomization, missing data, outcome measurement, etc.) as well as by an overall risk of bias assessment by outcome based on the Cochrane Risk of Bias 2 (RoB 2) tool (9). In addition, a variety of RCT and population characteristics are being extracted from each trial report including baseline severity, location, type of funding, presence of conflict of interest, and many others.

The COVID-NMA database is separated in two distinct databases which are imported and are continuously updated in *metaCOVID*: one containing data on all possible treatments against COVID-19 and one on COVID-19 vaccines. In this way, the application always uses the most up-to-date available databases for treatments and vaccines. As of November 2022, the treatment database included 493 randomized controlled trials (RCTs) and 522 pharmacologic and non-pharmacologic interventions forming in total 322 pairwise comparisons. Trial populations are separated into hospitalized patients and outpatients. We focus on binary and time-to-event outcomes as well as on both efficacy and safety (**Table 1**). With respect to vaccines, *metaCOVID* contains all available comparisons between a vaccine and placebo or no vaccination; up to November 2022 data from 117 RCTs on 55 different vaccines were available. The available pairwise comparisons are organized according to the platform type (e.g. RNA based vaccines, non-replicating viral vector vaccines, etc.). Given that the aim of *metaCOVID* is to perform data synthesis, it only uses data from RCTs to reassure that the setting of the studies will be similar to some degree. Data from observational studies for COVID-19 vaccines are presented in the COVID-NMA platform.

2.2 Data synthesis with metaCOVID

The application performs a meta-analysis of all the available RCTs based on the selection of the user regarding the comparison and the outcome of interest. For vaccines, where few RCTs and only comparisons between vaccines and controls are available, it is sufficient to specify the type of vaccine under consideration and to obtain the results for all vaccines under this category. Intervention effects are estimated using four different effect measures depending on the research question and the outcome of interest. Specifically, in our primary analysis, for treatments, risk ratios (RR) are used for binary outcomes and hazard ratios (HR) for time-to-event outcomes and for vaccines, vaccine efficacy (VE) is used for efficacy outcomes and RR for safety outcomes with $VE = (1 - RR) \times 100$. In the latter, RR may be the risk ratio or the rate ratio. All background computations are based on the R packages `metafor` and `meta` (10,11). The source code of our application is open access through GitHub at https://github.com/TEvrenoglou/metaCovid

By default, *metaCOVID* uses the inverse-variance approach and a random-effects model where the between-study variance τ^2 is estimated with the restricted maximum likelihood method (REML) (12–14). This is because COVID-19 trials are generally heterogeneous in several characteristics and this heterogeneity needs to be taken into account in the synthesis model. Although RCTs with zero events in both intervention arms are excluded from the analysis, they are included in the forest plots for completeness and transparency. For RCTs with zero events in one arm only we use the 0.5 continuity correction (12). Of course, the user can optionally change the default settings and select other estimation methods.

For example, when the number of available RCTs is small, the REML estimator for heterogeneity may underestimate the true between-study variance (13). The REML method may also rarely fail to provide an estimate of the heterogeneity due to lack of convergence of the iterative process that it is based on. Therefore, within *metaCOVID* we provide as an option alternative estimators: the DerSimonian-Laird estimator (15), the Paul-Mandel estimator (16), the Sidik-Jonkman estimator (17), the estimator that is based on empirical Bayes inference (18), and finally the maximum likelihood estimator (19). The different estimators may result in similar or different τ^2 estimates and, thus, it is useful to explore whether such differences impact the summary effect size estimated from the meta-analysis. The option of using a common-effect model assuming no heterogeneity across the RCTs of the meta-analysis is also available and may be used to monitor any differences in the summary results between the two models. Overall, the users should be careful when interpreting meta-analyses with only 2 or 3 RCTs since estimation of heterogeneity is always challenging in such small datasets (13,20).

2.3 Investigating heterogeneity within metaCOVID

Every forest plot produced by *metaCOVID* contains information on the τ^2 estimate, the (two-sided) *P*-value of the *Q*-test for homogeneity, the I^2 index – here is obtained as $I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2+\hat{s}^2} \times 100\%$ with $\hat{s}^2$ the 'overall' estimated within study variance) (21) – that gives the percentage of the total variation that is due to heterogeneity (22) , and the prediction interval (22). The latter represents the 95% interval within which the effect size of a future study is expected to lie. These

aid the user to infer on the presence or absence of important heterogeneity across the RCTs. When important heterogeneity is suspected, potential explanations need to be investigated. Typically, pre-specified subgroup analyses or meta-regressions are used to explore the impact of certain characteristics on the results. In *metaCOVID*, we only provide subgroup analyses since for most comparisons the number of available RCTs is not sufficient to perform meta-regression.

For treatments, a pre-selected subgroup analysis based on the baseline severity of the RCT participants is conducted by default for every comparison. This reflects the assumption that certain treatments might be effective for participants with severe and/or critical disease but not for those with a milder disease. We have already seen that this is a reasonable assumption at the trial level with the example of corticosteroids in RECOVERY (23). Additional optional subgroup analyses provided within the application are the type of the control intervention (standard care or placebo), the location of the study, the type of funding (public, mixed-private), and the presence or not of conflict of interest. Presentation of results without any subgroup is also possible. The latter might be particularly useful, for instance, when every severity subgroup has one RCT only and the sub-diamonds coincide with the RCT results. Every subgroup analysis is accompanied by the test for subgroup differences. This χ^2 test is similar to the Q-test but tests the assumption of homogeneity across the subgroups (24).

2.4 Assessing the robustness of results from metaCOVID

Sensitivity analyses show how robust the results remain when excluding certain RCTs which are considered suspicious for being substantially different (methodologically or clinically) from the other RCTs. A typical example here is the exclusion of RCTs rated being at high risk of bias (RoB). In *metaCOVID*, we go one step further and on top of the exclusion of high RoB RCTs we also allow the exclusion of RCTs with 'some concerns' for RoB. Overall, RoB is considered a key component within the application and, thus, the RoB assessments for all domains are presented along with the numerical results in our forest plots. However, here it should be noted that in *metaCOVID* the exclusion of trials is done only according to their overall RoB status. This is because in such an urgent situation it is very likely to compromise the quality of the studies in order to give

priority to the rapidity. In addition, we allow the users to exclude the preprints and obtain the results only from the RCTs published in medical journals. This sensitivity analysis aims to reflect and explore the concerns from some researchers and guideline developers with respect to the reliability of preprints.

For treatments, we further allow for a sensitivity analysis related to the assumption about missing outcome data. In our primary analysis, we follow the conservative approach of the intention-to-treat principle and we use for every intervention arm the number of participants randomized. Hence, participants for whom the outcome is missing are included in the analysis as non-events. It is important to highlight that in the presence of missing outcome data any approach relies on untestable assumptions and it is prone to bias if these assumptions are not valid (25,26). Consequently, the option of an available case analysis, where participants without the outcome are ignored, is also provided. The application further allows the calculation of the confidence interval through the Hartung-Knapp (HK) method (27), which has been found previously to result into better type I errors compared to the DL method (28). However, this method may yield very wide and non-informative confidence intervals in the presence of only 2-3 studies (29).

Finally, given that for some outcomes such as mortality the events are generally rare, the use of the inverse-variance model might not be optimal as results can be particularly imprecise or even biased (30). Other more appropriate models for rare events are available in the literature and should be used to evaluate the sensitivity of the results to the model choice. Within *metaCOVID* we provide three alternative approaches: the Mantel-Haenszel (MH) method, the Peto method, and the recently introduced penalized-likelihood meta-analysis (PL-MA) (31–33). The key advantage of these approaches is that they avoid the normal approximations that introduce bias in the presence of rare events and they model directly the arm-level data of the RCTs. Apart from less biased, these approaches usually provide also more precise results. In contrast to all other approaches, the PL-MA has the additional advantage that it allows incorporating in the analysis the double-zero RCTs without any continuity correction. This is because it uses prior information for the probability of an event through the Jeffrey's invariant prior which was originally proposed by Firth (33,34). However, the users should bear in mind that the Peto method and the PL-MA

only provide odds ratio (OR) estimates; hence the comparison of the results between this sensitivity analysis and the primary analysis might not be straightforward.

3 Results

We illustrate through different examples from the COVID-NMA database how *metaCOVID* can assist clinicians and guideline developers to explore the impact of different characteristics and modelling approaches on the results of meta-analysis. We chose examples involving well-known treatments and vaccines as well as several studies. For treatments, we use throughout RCTs on hospitalized adult patients for the outcome of all-cause mortality at 28 days as it is probably the most objective outcome. For vaccines, we use an example for the outcome of any adverse events. A tutorial on the use of the application can be found in our Supplementary material.

Figure 1 shows the results for the comparison convalescent plasma compared to standard care or placebo for all-cause mortality. There are 29 RCTs available with hospitalized patients and one of them includes only patients with mild disease. The summary RR suggests that there is probably no difference in efficacy between the experimental and the control intervention for this outcome. We can see quickly that this applies to mild or mixed in terms of baseline severity trial participants and that there is only one very small RCT at high RoB that has minimal impact (0.61% weight) on the results. Hence there are no concerns about the reliability of the results regarding RoB. However, the graph suggests some heterogeneity in the results of the studies that it is not reflected in the estimated $\hat{\tau}^2 = 0$. To explore this unexpected finding, we used also the other available estimators in *metaCOVID*. According to **Table 2**, the choice of the heterogeneity estimator has an impact on the results in this case. Specifically, all other estimators except maximum-likelihood suggest that there is some amount of heterogeneity across these RCTs with Sidnik-Jonkman suggesting that there might be even important heterogeneity ($\hat{\tau}^2 = 0.15$, $I^2 = 84.5\%$, $95\%\,\text{prediction interval} = [0.39, 1.84]$). **Table 2** also shows how the different heterogeneity estimators affect the point estimate of the summary RR and the lower limit of the confidence interval (CI). This sensitivity of the results to the choice of the estimator for heterogeneity should be taken into account when interpreting the results.

Figure 2 focuses on the comparison between hydroxychloroquine and control for all-cause mortality and presents a subgroup analysis where the RCTs have been separated based on the control intervention (standard care or placebo). Interestingly, hydroxychloroquine seems to be clearly worse than standard care but results are more uncertain when it is compared to placebo. Although the test for subgroup differences suggest that there is not a statistically significant difference (p=0.24), providing potential explanations for this difference is still important. Obvious candidate characteristics seem to be the use of preprints or RCTs at high risk of bias; nevertheless, excluding such RCTs and/or preprints does not impact on the results. We can see, though, that there are two large RCTs in the standard care group that affect a lot the results, while at the placebo group we miss RCTs with so large sample size. In such a case, a funnel plot could be further drawn to assist the conclusions.

Mortality is a generally rare outcome in the context of COVID-19. We conducted a sensitivity analysis for the comparison between favipiravir and standard care or placebo seven RCTs are available but only three have some events observed. This is a situation where the inverse-variance approach might be biased and other methods should be preferred. **Table 3** shows that the three approaches (MH, Peto, and PL-MA) give very similar OR estimates and 95% CI, whereas the inverse-variance approach yields substantially smaller OR estimate and different 95% CI yet to the same direction.

Figure 3 shows the comparisons between non-replicating viral vector vaccines and control for the incident of any adverse events. As expected, the vaccines cause more adverse events than the control interventions. All RCTs here but two are at low RoB, while these three RCTs with some concerns tend to give more favorable results for the two corresponding vaccines (AstraZeneca and Janssen) than other RCTs. Excluding these two RCTs has a small impact on the summary RR estimate for AstraZeneca but a large impact for Janssen that moves from 1.50 (0.92, 2.47) to 2.31 (1.80, 2.97). This change, though, should be interpreted with caution as, according to **Figure 3**, the study that was excluded for the Janssen vaccine is a very large RCT which might be more appropriate to estimate correctly the adverse events of the vaccine than a small RCT. Interestingly, the only study suggesting less adverse events for the vaccine (AstraZeneca) than the control involves two control interventions (i.e. another vaccine or placebo); as expected it

might not be appropriate for this outcome to combine RCTs using active controls with RCTs using placebo as control.

4 Discussion

In this paper, we present a new online application, called *metaCOVID*, that offers a user-friendly environment for performing living meta-analyses of RCTs for COVID-19 treatments and vaccines. This application is the core tool for updating regularly the analyses of the COVID-NMA platform (35,36), but it is also used by external organizations, such as Cochrane South Africa and the UK National Institute for Health and Care Excellence (NICE), to supplement their process for forming recommendations. This application offers open access to the most-up-to date database of COVID-19 RCTs for researchers, clinicians, or guideline developers interested to perform amendable meta-analyses and explore the impact of certain characteristics on the results. Users can download and freely use the outputs from the application given that they recognize *metaCOVID* and the COVID-NMA platform.

A key characteristic of *metaCOVID* is that it gives the opportunity to the users to undertake fast multiple analyses and present complex data structures without requiring any technical and software knowledge. At the same time, it preserves from arbitrary selections of variables for performing subgroup and sensitivity analyses by following a pre-defined protocol (1,2). The living approach implemented in *metaCOVID* does not apply only to the continuous incorporation of new RCTs but designates all aspects of the systematic review process; namely all considerations regarding data extraction, risk of bias assessments, data synthesis, and appraisal of results are re-evaluated regularly based on the evolution of the data and the knowledge for the disease. Any proposed changes need to be accepted by the steering committee of the COVID-NMA. For instance, one important change that was decided when concerns were raised by methodologists of the steering committee and after several discussions with clinical experts was to resign from the plan of a network meta-analysis of all the COVID-19 treatments and focus on smaller network meta-analyses for treatments with similar characteristics. That was decided to avoid obtaining misleading and biased results from synthesizing highly heterogeneous interventions and

populations and violating the assumptions. Therefore, *metaCOVID* is currently restricted to pairwise meta-analysis only.

The present version of *metaCOVID* is restricted to the use of the COVID-NMA database. This certainly limits the applicability of the tool to a different setting but future versions will allow the users to upload their data. Also, so far, the databases imported in metaCOVID cannot be downloaded; this will be possible, though, in future updates of application. To avoid large amounts of heterogeneity and misleading conclusions, we have further restricted the application to meta-analyses of RCTs. In this way, possibly important information from observation studies is disregarded. An additional limitation of our application is related to the use of only pre-defined subgroup and sensitivity analyses in the COVID-NMA protocol. However, this reassures that all selected variables have been selected based on scientific rationale and relevant clinical expertise. Finally, as with all similar tools, the ease that *metaCOVID* provides for performing complex analyses can also be a drawback if the application is used without cautiousness. Users should be aware that interpretation of meta-analysis results requires understanding of the data and the clinical setting, considerations on the credibility of the evidence, and investigation of all possible sources of within- and across-study biases (7,37). Hence, although the application provides a user-friendly environment, it cannot guarantee the proper interpretation of the findings by the users.

Our application serves as a pilot version of a generalized application that could be tailored for living meta-analyses of any condition where the users will be able to upload their own data. Although, here the application is linked to the COVID-NMA platform, it can also be seen as a standalone tool since its use does not require any knowledge of the material available in the platform. All living evidence synthesis projects are highly resource- and time-demanding as they involve large amounts of data that need to be continuously updated and analyzed. Rapid and efficient automation tools are required for the sustainability of any living systematic review that aims to have a very short delay (i.e. a few weeks) between every update as well as to allow for immediate dissemination and access to the findings. There is a general tendency in evidence synthesis towards long-term projects and broader reviews covering several research questions; such projects are impossible to move forward without tailored software facilitating and

accelerating the data manipulation and the analysis. In addition, online platforms and software applications promote open-access research provided that they are used carefully and results are interpreted with a global view of all the available information.

Acknowledgements: The authors would like to thank the COVID-NMA consortium and particularly the data extraction team for providing the data used in *metaCOVID* as well as for their useful feedback on the application.

Data availability statement: The data are publicly available at: https://covid-nma.com/

Declarations of interest: none.

Highlights

What is already known:

- Living systematic reviews and meta-analyses are becoming increasingly popular.
- Existing living systematic reviews are limited by lack of automation tools to allow fast production and dissemination of results and direct access to the data

What this adds:

- *metaCOVID* is an online tool that facilitates the exploration and analysis of complex data structures, such as the COVID-NMA database, without requiring technical and programming skills.
- This application provides free access to the most up-to-date database for COVID-19 trials for treatment and vaccine comparisons and can assist guideline developers to form timely recommendations.
- Using this open-access application, different stakeholders can rapidly investigate the impact of different characteristics on the results and easily produce comprehensive graphical displays.

Potential impact for *Research Synthesis Methods* readers outside the authors' field

- This pilot application forms the basis for creating fast and efficient automation tools that accelerate the analyses and the dissemination of the findings for future living evidence synthesis projects.
- Freely available software tools, like *metaCOVID,* are indispensable tools in the era of open science but should be accompanied by careful and thoughtful interpretation of the outputs.

References

1. Boutron I, Chaimani A, Devane D, Meerpohl J, Rada G, Hróbjartsson A, et al. Interventions for the prevention and treatment of COVID-19: a living mapping of research and living network meta-analysis. Cochrane Database of Systematic Reviews [Internet]. 2020;(11). Available from: https://doi.org//10.1002/14651858.CD013769

2. Boutron I, Chaimani A, Devane D, Meerpohl J, Rada G, Hróbjartsson A, et al. Interventions for the treatment of COVID-19: a living network meta-analysis. Cochrane Database of Systematic

Reviews [Internet]. 2020;(11). Available from: https://doi.org//10.1002/14651858.CD013770

3. Boutron I, Chaimani A, Meerpohl JJ, Hróbjartsson A, Devane D, Rada G, et al. The COVID-NMA Project: Building an Evidence Ecosystem for the COVID-19 Pandemic. Ann Intern Med. 2020 Dec 15;173(12):1015–7.

4. Siemieniuk RA, Bartoszko JJ, Ge L, Zeraatkar D, Izcovich A, Kum E, et al. Drug treatments for covid-19: living systematic review and network meta-analysis. Vol. 370, BMJ. 2020. p. m2980.

5. Juul S, Nielsen N, Bentzer P, Veroniki AA, Thabane L, Linder A, et al. Interventions for treatment of COVID-19: a protocol for a living systematic review with network meta-analysis including individual patient data (The LIVING Project). Syst Rev. 2020 May 9;9(1):108.

6. Malin JJ, Spinner CD, Janssens U, Welte T, Weber-Carstens S, Schälte G, et al. Key summary of German national treatment guidance for hospitalized COVID-19 patients: Key pharmacologic recommendations from a national German living guideline using an Evidence to Decision Framework (last updated 17.05.2021). Infection. 2022 Feb;50(1):93–106.

7. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ [Internet]. 2015;351. Available from: https://www.bmj.com/content/351/bmj.h5651

8. Verdugo-Paiva F, Vergara C, Ávila C, Castro-Guevara JA, Cid J, Contreras V, et al. COVID-19 Living Overview of Evidence repository is highly comprehensive and can be used as a single source for COVID-19 studies. Journal of Clinical Epidemiology. 2022;149:195–202.

9. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug 28;366:l4898.

10. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. 2010th-08–05 ed. Vol. 36, Journal of Statistical Software. 2010. p. 48.

11. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health. 2019 Nov;22(4):153–60.

12. Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta-analyses. Cochrane handbook for systematic reviews of interventions. 2019;241–84.

13. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods. 2016;7(1):55–79.

14. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. Journal of Educational and Behavioral Statistics. 2005;30(3):261–93.

15. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986;7(3):177–88.

16. Paule RC, Mandel J. Consensus values and weighting factors. Journal of Research of the National Bureau of Standards. 1982;87(5):377–85.

17. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005;54(2):367–84.

18. Morris CN. Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association. 1983;78(381):47–55.

19. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med. 1996 Mar 30;15(6):619–29.

20. Petropoulou M, Mavridis D. A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study. Statistics in medicine. 2017;36(27):4266–80.

21. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in medicine. 2002;21(11):1539–58.

22. Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Research synthesis methods. 2017;8(1):5–18.

23. Dexamethasone in Hospitalized Patients with Covid-19. Vol. 384, New England Journal of Medicine. 2020. p. 693–704.

24. Borenstein M, Higgins JP. Meta-analysis and subgroups. Prevention science. 2013;14(2):134–43.

25. Mavridis D, White IR. Dealing with missing outcome data in meta-analysis. Research Synthesis Methods. 2020;11(1):2–13.

26. Mavridis D, Chaimani A, Efthimiou O, Leucht S, Salanti G. Addressing missing outcome data in meta-analysis. Evidence-based mental health. 2014;17(3):85–9.

27. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Statistics in Medicine. 2003;22(17):2693–710.

28. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Medical Research Methodology. 2014 Feb 18;14(1):25.

29. Bender R, Friede T, Koch A, Kuss O, Schlattmann P, Schwarzer G, et al. Methods for evidence synthesis in the case of very few studies. Research Synthesis Methods. 2018;9(3):382–92.

30. Efthimiou O. Practical guide to the meta-analysis of rare events. Evid Based Ment Health. 2018 May 1;21(2):72.

31. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the national cancer institute. 1959;22(4):719–48.

32. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Progress in cardiovascular diseases. 1985;27(5):335–71.

33. Evrenoglou T, White IR, Afach S, Mavridis D, Chaimani A. Network meta-analysis of rare events using penalized likelihood regression. Statistics in Medicine. 2022;41(26):5203–19.

34. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;27–38.

35. Ghosn L, Chaimani A, Evrenoglou T, Davidson M, Graña C, Schmucker C, et al. Interleukin-6 blocking agents for treating COVID-19: a living systematic review. Cochrane Database of Systematic Reviews [Internet]. 2021;(3). Available from: https://doi.org//10.1002/14651858.CD013881

36. Davidson M, Menon, S, Chaimani, A, Evrenoglou, T, Ghosn, L, Graña, C, Henschke, N, Cogo, E, Villanueva, G, Ferrand, G, Riveros, C, Bonnet, H, Kapp, P, Moran, C, Devane, D, Meerpohl, JJ, Rada, G, Hróbjartsson, A, Grasselli, G, Tovey, D, Ravaud, P, Boutron I. Interleukin-1 blocking agents for treating COVID-19. Cochrane Database of Systematic Reviews [Internet]. 2022;(1). Available from: https://doi.org//10.1002/14651858.CD015308

37. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. Canadian Medical Association journal. 2020 Aug 10;192(32):E901–6.

Tables

Table 1: List of all the outcomes and corresponding effect sizes incorporated in *metaCOVID*.

Outcome	Effect size	Research question
All-cause mortality	Risk ratio	COVID-19 treatments/vaccines
Clinical improvement	Risk ratio	COVID-19 treatments
WHO Clinical Progression Score level 7 or above	Risk ratio	COVID-19 treatments
Viral negative conversion	Risk ratio	COVID-19 treatments
Incidence of any adverse events	Risk ratio	COVID-19 treatments/vaccines
Incidence of serious adverse events	Risk ratio	COVID-19 treatments/vaccines
Hospitalization or death	Risk ratio	COVID-19 treatments
Time to clinical improvement	Hazard ratio	COVID-19 treatments
Time to death	Hazard ratio	COVID-19 treatments
Time to WHO Clinical Progression Score level 7 or above	Hazard ratio	COVID-19 treatments
Time to viral negative conversion	Hazard ratio	COVID-19 treatments
Confirmed severe or critical disease due to Covid-19	Vaccine efficacy	COVID-19 vaccines
Symptomatic Covid-19	Vaccine efficacy	COVID-19 vaccines
SARS-CoV-2 infection	Vaccine efficacy	COVID-19 vaccines
Systemic adverse events	Risk ratio	COVID-19 vaccines

Table 2: Results for heterogeneity across the different estimators provided by *metaCOVID*. The comparison of interest is convalescent plasma versus standard care or placebo and the outcome of interest is all-cause mortality at day 28.

Estimation method	τ^2	I^2	95% Prediction Interval	Pooled RR [95% CI]
REML	0	0%	[0.93, 1.03]	0.98 [0.93. 1.03]
Maximum likelihood	0	0%	[0.93, 1.03]	0.98 [0.93. 1.03]
DerSimonian-Laird	0.01	17.7%	[0.79, 1.14]	0.95 [0.86, 1.04]
Sidik Jonkman	0.14	82.1%	[0.39, 1.84]	0.85 [0.70, 1.04]
Empirical Bayes	0.03	47.9%	[0.64, 1.31]	0.91 [0.80, 1.04]
Paule-Mandel	0.03	47.9%	[0.64, 1.31]	0.91 [0.80, 1.04]

Table 3: Sensitivity analysis for the comparison between favipiravir and standard care or placebo using alternative methods which are suitable for rare events. The outcome of interest is all-cause mortality at day 28.

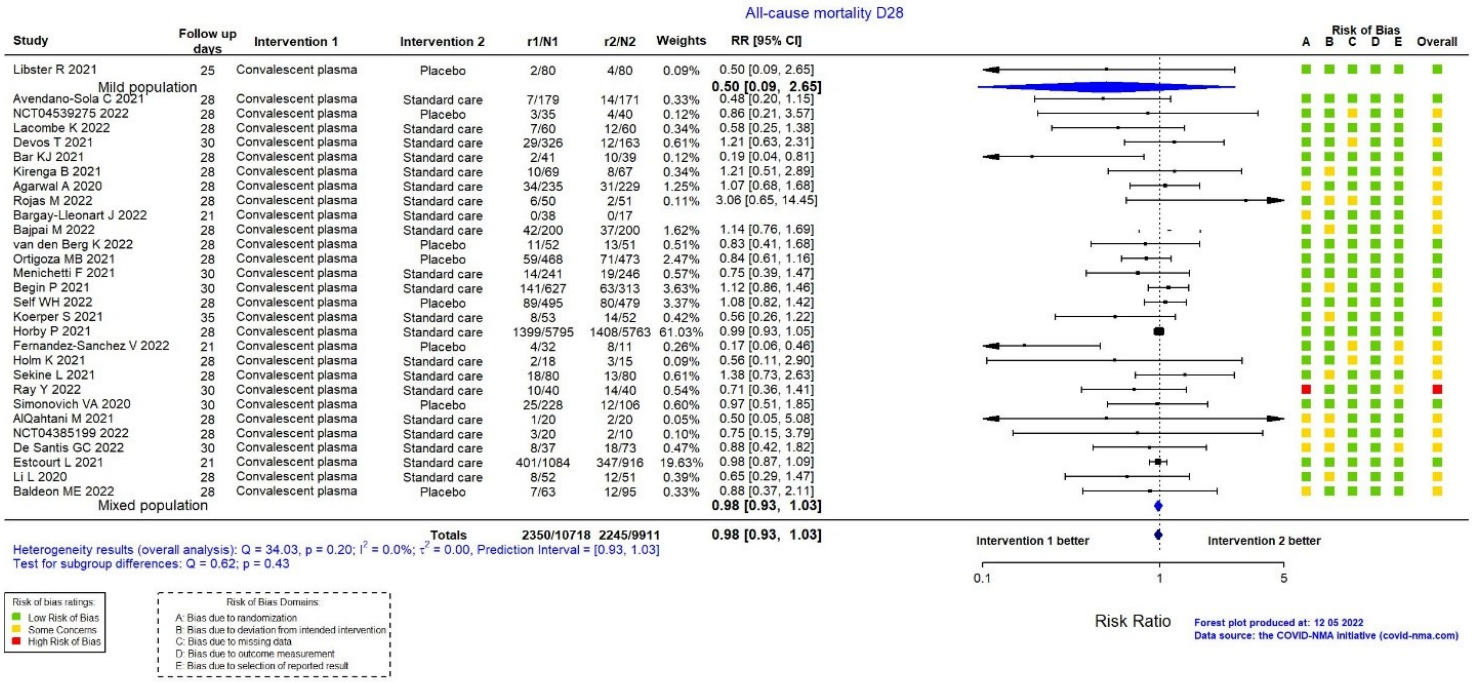
Meta-analysis model	Pooled OR [95% CI]
Inverse-variance	1.34 [0.67, 2.65]
Mantel-Haenszel	1.75 [0.83 3.65]
Peto	1.72 [0.84,3.51]
Logistic regression with penalized likelihood	1.71 [0.86, 3.40]

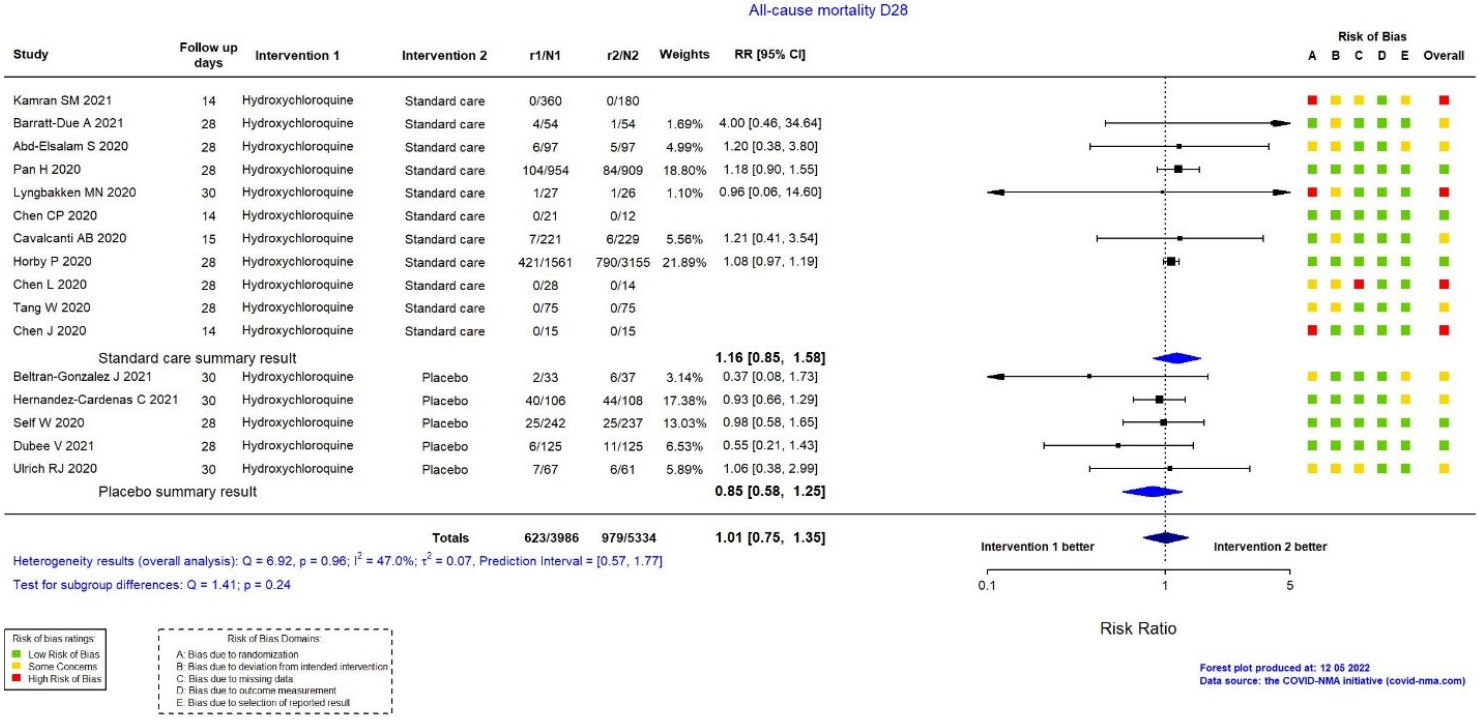
Figures

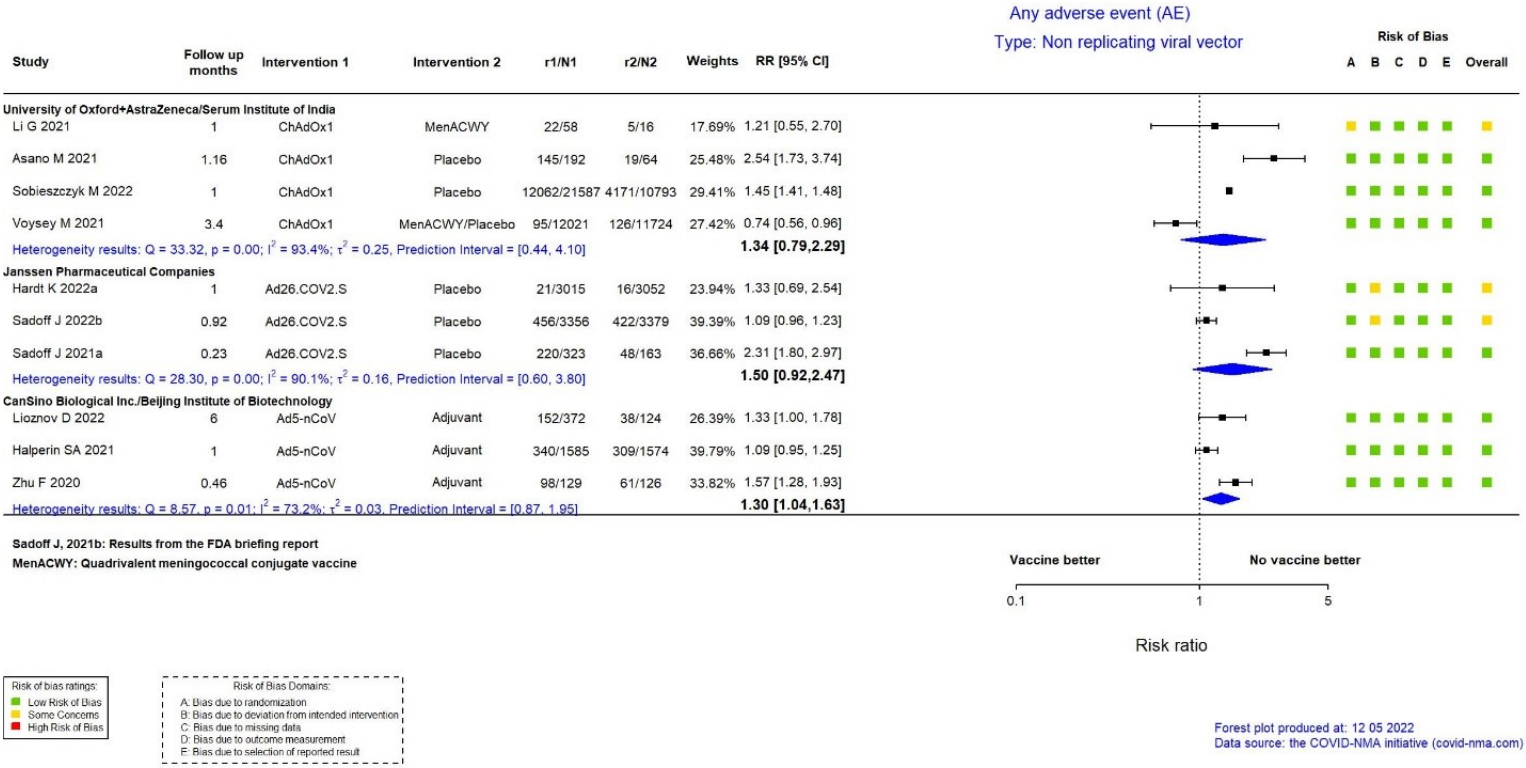
Figure 1: Forest plot produced from *metaCOVID* for the comparison between convalescent plasma versus standard care or placebo for hospitalized patients.

Figure 2: Forest plot produced from *metaCOVID* for the comparison between hydroxychloroquine versus standard care or placebo for hospitalized patients.

Figure 3: Forest plot produced from *metaCOVID* for the comparisons between non-replicating viral vector vaccines and control for the incident of any adverse events.


All-cause mortality D28
Study
Follow up days
Intervention 1
Intervention 2
r1/N1
r2/N2
Weights
RR [95% CI]
Risk of Bias
A B C D E Overall
Libster R 2021 25 Convalescent plasma Placebo 2/80 4/80 0.09% 0.50 [0.09, 2.65]
Mild population 0.50 [0.09, 2.65]
Avendano-Sola C 2021 28 Convalescent plasma Standard care 7/179 14/171 0.33% 0.48 [0.20, 1.15]
NCT04539275 2022 28 Convalescent plasma Placebo 3/35 4/40 0.12% 0.86 [0.21, 3.57]
Lacombe K 2022 28 Convalescent plasma Standard care 7/60 12/60 0.34% 0.58 [0.25, 1.38]
Devos T 2021 30 Convalescent plasma Standard care 29/326 12/163 0.61% 1.21 [0.63, 2.31]
Bar KJ 2021 28 Convalescent plasma Standard care 2/41 10/39 0.12% 0.19 [0.04, 0.81]
Kirenga B 2021 28 Convalescent plasma Standard care 10/69 8/67 0.34% 1.21 [0.51, 2.89]
Agarwal A 2020 28 Convalescent plasma Standard care 34/235 31/229 1.25% 1.07 [0.68, 1.68]
Rojas M 2022 28 Convalescent plasma Standard care 6/50 2/51 0.11% 3.06 [0.65, 14.45]
Bargay-Lleonart J 2022 21 Convalescent plasma Standard care 0/38 0/17
Bajpai M 2022 28 Convalescent plasma Standard care 42/200 37/200 1.62% 1.14 [0.76, 1.69]
van den Berg K 2022 28 Convalescent plasma Placebo 11/52 13/51 0.51% 0.83 [0.41, 1.68]
Ortigoza MB 2021 28 Convalescent plasma Placebo 59/468 71/473 2.47% 0.84 [0.61, 1.16]
Menichetti F 2021 30 Convalescent plasma Standard care 14/241 19/246 0.57% 0.75 [0.39, 1.47]
Begin P 2021 30 Convalescent plasma Standard care 141/627 63/313 3.63% 1.12 [0.86, 1.46]
Self WH 2022 28 Convalescent plasma Placebo 89/495 80/479 3.37% 1.08 [0.82, 1.42]
Koerper S 2021 35 Convalescent plasma Standard care 8/53 14/52 0.42% 0.56 [0.26, 1.22]
Horby P 2021 28 Convalescent plasma Standard care 1399/5795 1408/5763 61.03% 0.99 [0.93, 1.05]
Fernandez-Sanchez V 2022 21 Convalescent plasma Placebo 4/32 8/11 0.26% 0.17 [0.06, 0.46]
Holm K 2021 28 Convalescent plasma Standard care 2/18 3/15 0.09% 0.56 [0.11, 2.90]
Sekine L 2021 28 Convalescent plasma Standard care 18/80 13/80 0.61% 1.38 [0.73, 2.63]
Ray Y 2022 30 Convalescent plasma Standard care 10/40 14/40 0.54% 0.71 [0.36, 1.41]
Simonovich VA 2020 30 Convalescent plasma Placebo 25/228 12/106 0.60% 0.97 [0.51, 1.85]
AlQahtani M 2021 28 Convalescent plasma Standard care 1/20 2/20 0.05% 0.50 [0.05, 5.08]
NCT04385199 2022 28 Convalescent plasma Standard care 3/20 2/10 0.10% 0.75 [0.15, 3.79]
De Santis GC 2022 30 Convalescent plasma Standard care 8/37 18/73 0.47% 0.88 [0.42, 1.82]
Estcourt L 2021 21 Convalescent plasma Standard care 401/1084 347/916 19.63% 0.98 [0.87, 1.09]
Li L 2020 28 Convalescent plasma Standard care 8/52 12/51 0.39% 0.65 [0.29, 1.47]
Baldeon ME 2022 28 Convalescent plasma Placebo 7/63 12/95 0.33% 0.88 [0.37, 2.11]
Mixed population 0.98 [0.93, 1.03]
Totals 2350/10718 2245/9911 0.98 [0.93, 1.03]
Heterogeneity results (overall analysis): Q = 34.03, p = 0.20; I² = 0.0%; τ² = 0.00, Prediction Interval = [0.93, 1.03]
Test for subgroup differences: Q = 0.62; p = 0.43
Intervention 1 better
Intervention 2 better
0.1
1
5
Risk Ratio
Forest plot produced at: 12 05 2022
Data source: the COVID-NMA initiative (covid-nma.com)
Risk of bias ratings:
Low Risk of Bias
Some Concerns
High Risk of Bias
Risk of Bias Domains:
A: Bias due to randomization
B: Bias due to deviation from intended intervention
C: Bias due to missing data
D: Bias due to outcome measurement
E: Bias due to selection of reported result


All-cause mortality D28
Study
Follow up days
Intervention 1
Intervention 2
r1/N1
r2/N2
Weights
RR [95% CI]
Risk of Bias
A B C D E Overall
Kamran SM 2021 14 Hydroxychloroquine Standard care 0/360 0/180
Barratt-Due A 2021 28 Hydroxychloroquine Standard care 4/54 1/54 1.69% 4.00 [0.46, 34.64]
Abd-Elsalam S 2020 28 Hydroxychloroquine Standard care 6/97 5/97 4.99% 1.20 [0.38, 3.80]
Pan H 2020 28 Hydroxychloroquine Standard care 104/954 84/909 18.80% 1.18 [0.90, 1.55]
Lyngbakken MN 2020 30 Hydroxychloroquine Standard care 1/27 1/26 1.10% 0.96 [0.06, 14.60]
Chen CP 2020 14 Hydroxychloroquine Standard care 0/21 0/12
Cavalcanti AB 2020 15 Hydroxychloroquine Standard care 7/221 6/229 5.56% 1.21 [0.41, 3.54]
Horby P 2020 28 Hydroxychloroquine Standard care 421/1561 790/3155 21.89% 1.08 [0.97, 1.19]
Chen L 2020 28 Hydroxychloroquine Standard care 0/28 0/14
Tang W 2020 28 Hydroxychloroquine Standard care 0/75 0/75
Chen J 2020 14 Hydroxychloroquine Standard care 0/15 0/15
Standard care summary result 1.16 [0.85, 1.58]
Beltran-Gonzalez J 2021 30 Hydroxychloroquine Placebo 2/33 6/37 3.14% 0.37 [0.08, 1.73]
Hernandez-Cardenas C 2021 30 Hydroxychloroquine Placebo 40/106 44/108 17.38% 0.93 [0.66, 1.29]
Self W 2020 28 Hydroxychloroquine Placebo 25/242 25/237 13.03% 0.98 [0.58, 1.65]
Dubee V 2021 28 Hydroxychloroquine Placebo 6/125 11/125 6.53% 0.55 [0.21, 1.43]
Ulrich RJ 2020 30 Hydroxychloroquine Placebo 7/67 6/61 5.89% 1.06 [0.38, 2.99]
Placebo summary result 0.85 [0.58, 1.25]
Totals 623/3986 979/5334 1.01 [0.75, 1.35]
Heterogeneity results (overall analysis): Q = 6.92, p = 0.96; I² = 47.0%; τ² = 0.07, Prediction Interval = [0.57, 1.77]
Test for subgroup differences: Q = 1.41; p = 0.24
Intervention 1 better
Intervention 2 better
0.1
1
5
Risk Ratio
Risk of bias ratings:
Low Risk of Bias
Some Concerns
High Risk of Bias
Risk of Bias Domains:
A: Bias due to randomization
B: Bias due to deviation from intended intervention
C: Bias due to missing data
D: Bias due to outcome measurement
E: Bias due to selection of reported result
Forest plot produced at: 12 05 2022
Data source: the COVID-NMA initiative (covid-nma.com)


Any adverse event (AE)
Type: Non replicating viral vector
Study
Follow up months
Intervention 1
Intervention 2
r1/N1
r2/N2
Weights
RR [95% CI]
Risk of Bias
A B C D E Overall
University of Oxford+AstraZeneca/Serum Institute of India
Li G 2021 1 ChAdOx1 MenACWY 22/58 5/16 17.69% 1.21 [0.55, 2.70]
Asano M 2021 1.16 ChAdOx1 Placebo 145/192 19/64 25.48% 2.54 [1.73, 3.74]
Sobieszczyk M 2022 1 ChAdOx1 Placebo 12062/21587 4171/10793 29.41% 1.45 [1.41, 1.48]
Voysey M 2021 3.4 ChAdOx1 MenACWY/Placebo 95/12021 126/11724 27.42% 0.74 [0.56, 0.96]
Heterogeneity results: Q = 33.32, p = 0.00; I² = 93.4%; τ² = 0.25, Prediction Interval = [0.44, 4.10]
1.34 [0.79,2.29]
Janssen Pharmaceutical Companies
Hardt K 2022a 1 Ad26.COV2.S Placebo 21/3015 16/3052 23.94% 1.33 [0.69, 2.54]
Sadoff J 2022b 0.92 Ad26.COV2.S Placebo 456/3356 422/3379 39.39% 1.09 [0.96, 1.23]
Sadoff J 2021a 0.23 Ad26.COV2.S Placebo 220/323 48/163 36.66% 2.31 [1.80, 2.97]
Heterogeneity results: Q = 28.30, p = 0.00; I² = 90.1%; τ² = 0.16, Prediction Interval = [0.60, 3.80]
1.50 [0.92,2.47]
CanSino Biological Inc./Beijing Institute of Biotechnology
Lioznov D 2022 6 Ad5-nCoV Adjuvant 152/372 38/124 26.39% 1.33 [1.00, 1.78]
Halperin SA 2021 1 Ad5-nCoV Adjuvant 340/1585 309/1574 39.79% 1.09 [0.95, 1.25]
Zhu F 2020 0.46 Ad5-nCoV Adjuvant 98/129 61/126 33.82% 1.57 [1.28, 1.93]
Heterogeneity results: Q = 8.57, p = 0.01; I² = 73.2%; τ² = 0.03, Prediction Interval = [0.87, 1.95]
1.30 [1.04,1.63]
Sadoff J, 2021b: Results from the FDA briefing report
MenACWY: Quadrivalent meningococcal conjugate vaccine
Vaccine better
No vaccine better
0.1
1
5
Risk ratio
Risk of bias ratings:
Low Risk of Bias
Some Concerns
High Risk of Bias
Risk of Bias Domains:
A: Bias due to randomization
B: Bias due to deviation from intended intervention
C: Bias due to missing data
D: Bias due to outcome measurement
E: Bias due to selection of reported result
Forest plot produced at: 12 05 2022
Data source: the COVID-NMA initiative (covid-nma.com)

4. Part 3: Dealing with sparse networks in network meta-analysis

The validity of NMA estimates can be threatened in the presence of sparse networks[29,57,86]. Such cases refer to networks with limited direct comparisons and only few available studies to inform them. Due to the low amount of the available information the large sample approximation assumptions to which NMA models rely are invalid. As a consequence, the final NMA estimates are expected to be imprecise and potentially biased[29]. In addition, the estimation of the heterogeneity parameter can be challenging when only few studies are available[34,35]. A problematic estimation of the heterogeneity parameter can lead to important problems related to the confidence or credible intervals of the final NMA estimates as those are calculated based both on the within and the across-studies variance[32].

For sparse networks uncertainty can also exist in the formal evaluation of the underlying NMA assumptions[66,86]. The transitivity assumption requires balanced effect modifier distributions across all pairs of interventions which involve a common comparator. However, such an evaluation can be challenging if the network is sparse as every comparison is only informed by 2-3 studies. Any potential intransitivity in the network is expected to appear in the data as discrepancies between direct and indirect NMA estimates. This leads to the statistical manifestation of the transitivity assumption namely the consistency assumption. Within the frequentist NMA framework inconsistency checks usually rely on hypothesis testing methods which were defined by equations (1.9)-(1.11). For NMA models fitted in the Bayesian framework such checks rely on the comparison between the posterior densities of both direct and indirect estimates[63,65,78]. Both of those types of checks can be challenged in the presence of sparse networks. The hypothesis testing

approaches are underpowered while potential important differences between the posterior densities can be masked due to the large uncertainty that inevitably accompanies both direct and indirect estimates when only few studies are available[45].

The aim of the third project of this thesis is to deal with sparse networks in NMA. Specifically, sparse networks typically arise for 'sensitive' subgroups of the population which cannot be easily included in studies such as children, elder patients, or individuals with multimorbidity[88]. In such cases of sparsity, external evidence can be used to facilitate the estimation of the intervention effects[88,91]. In this project the emphasis is given on different alternatives for sharing information between two networks that pertain to different subgroups of the population: a target subgroup that forms a sparse network and a different subgroup that forms a 'dense' network.

Sharing of information between different subgroups of the population should always account for the differences in the underlying subgroup-specific intervention effects. In addition, the process of sharing information should also account for avoiding the dominance of the external information over the target data. To account for all the previously mentioned issues a two-stage Bayesian framework is suggested to facilitate the analysis of a sparse network of interventions while borrowing strength from a dense network.

At the first-stage a modified NMA model is used for the analysis only of the dense network. The use of this modified model aims to extrapolate the NMA estimates of the dense network to those of the sparse network of interventions. A scale parameter is added to the within-studies assumption in order to inflate the variance of the external data as those target a different population than the population of primary interest. A location parameter is added for the across-studies assumption. This parameter aims to shift the distribution of the relative effects in the dense network towards the distribution of relative effects in the sparse network. The extrapolated NMA estimates

obtained from the analysis of the dense network are then considered as applicable to the sparse network. The latter can only be achieved if a prior distribution is properly assigned for the location and scale parameters. In this project the construction of a prior distribution for the location parameter can be based either on a data based approach or on prior elicitation based on expert opinion. The choice of the prior distribution for the scale parameter can be based on the expert opinion regarding the importance of the clinical differences between the two populations.

At the second-stage the predictive distributions of the extrapolated NMA estimates are used as informative prior distributions for the analysis of the sparse network. Specifically, at this stage the analysis of the sparse network takes place by using the hierarchical NMA model fitted in the Bayesian framework. The key difference between this model and the standard NMA model is that in this case the informative prior distributions obtained at the first-stage are used as priors for the basic parameters while in the standard NMA approach non-informative priors are conventionally assigned for the basic parameters.

The use of the two-stage approach is illustrated through an example of two subgroups of psychotic patients: the target subgroup of children-adolescents (CA) that forms a rather sparse network and the subgroup of chronic adult patients with positive symptoms, referred as “general patients” (GP), that forms a dense network. At the first-stage the results of the dense network of GP are extrapolated to those of the CA using the modified NMA model. This model is fitted with the prior for the location parameter constructed once by using a data-based approach and once based on expert opinion. For the prior elicitation using expert’s opinion a questionnaire was circulated among experts with experience in the field of schizophrenia. The two-stage approach was compared with two other approaches. At first with a naïve synthesis model which assumes that GP and CA are equivalent sources of information and thus combines the two networks into a

single one. At second with the standard NMA approach which places non-informative priors for the basic parameters and thus does not make any use of external evidence.

The findings of the illustrative example suggest that the use of external evidence through the two-stage approach can facilitate the estimation of intervention effects in cases of sparse networks of interventions. Specifically, the use of informative priors was found to substantially improve the precision of the NMA estimates in comparison to those obtained from the standard NMA model with non-informative priors. Moreover, the unsuitability of the standard NMA model to handle sparse networks is also depicted through its comparison with the simple direct estimates. Those direct estimates found to be more precise than estimates obtained by the standard NMA model for almost every comparison. Finally, as expected the most precise estimates were obtained from the naïve synthesis model. However, this model does not account for any difference between the population of GP and CA and treats them as equivalent sources of information. This implies that the NMA estimates obtained from such a model are very likely to be dominated by the presence of the external data.

A detailed description regarding the new approach for analyzing sparse networks is given in the following manuscript: Evrenoglou, T, Metelli, S, Johannes-Schneider, T, Siafis, S, Leucht, S, Chaimani, A. Sharing information across patient subgroups to draw conclusions from sparse treatment networks [to be submitted until 30th of November at the ISCB special issue of the Biometrical Journal]

The corresponding supporting information for this article are given in Annex 3 of the thesis.

Sharing information across patient subgroups to draw conclusions from sparse treatment networks

Theodoros Evrenoglou[1], Silvia Metelli[1], Johannes-Schneider Thomas[2], Spyridon Siafis[2], Stefan Leucht[2], Anna Chaimani[1]

[1] *Université Paris Cité, Research Center of Epidemiology and Statistics (CRESS-U1153), INSERM, Paris, France*

[2] *Department of Psychiatry and Psychotherapy, School of Medicine, Technical University of Munich, Germany*

Abstract: Network meta-analysis (NMA) usually provides estimates of the relative effects with the highest possible precision. However, sparse networks with few available studies and limited direct evidence can arise, threatening the robustness and reliability of NMA estimates. This is often the case for 'sensitive' population subgroups for which recruitment is more challenging, such as children, elder or multimorbid patients. In these cases, the limited amount of available information can hamper the formal evaluation of the underlying NMA assumptions of transitivity and consistency. In addition, NMA estimates from sparse networks are expected to be biased and imprecise as they rely on large sample approximations which would be here invalid. We propose a Bayesian framework that allows to share information between two networks that pertain to different population subgroups. Specifically, we use the results from a subgroup with a lot of direct evidence (a dense network) to construct informative priors for the relative effects of the target subgroup (a sparse network). This is a two-stage approach where at the first stage we extrapolate the results of the dense network to those expected from the sparse network. This takes place by using a modified hierarchical NMA model where we add a location parameter that shifts the distribution of the relative effects to make them applicable to the target population. The location parameter is informed either through the data or using expert opinion. At the second stage, these extrapolated results are used as prior information for the sparse network. We illustrate our approach through a motivating example of psychiatric patients where the target subgroup of children-adolescents forms a sparse network and the subgroup of "general patients" forms a dense network. Our approach results in more precise and robust estimates of the relative effects which can give insight about the efficacy of antipsychotics for the target subgroup and adequately inform clinical practice.

Keywords: sharing information, mixed treatment comparisons, informative priors, sparse data

1. Introduction

Network meta-analysis (NMA) has become an essential tool for the comparative effectiveness research of healthcare interventions since it allows integrating all available information on a specific disease and usually provides estimates of relative effects with the highest possible precision[1,2]. However, networks of interventions with limited data that fail to provide useful conclusions may arise under certain situations. A recent empirical study found 92 (7,4%) published NMAs that included more treatments than the number of studies in a sample of 1236 NMAs of randomized controlled trials (RCTs) with at least four interventions[3]. These numbers suggest that the phenomenon of 'sparse' networks, namely networks with limited direct evidence in the form of few direct comparisons and/or few - 1 or 2 - studies per comparison, is not very rare in the literature. For example, this is often the case for 'sensitive' subgroups of the population that cannot be easily included in trials, such as children[4], elder patients, or individuals with multimorbidity.

Results from such networks of interventions are accompanied with substantial uncertainty not only in the estimation of the relative treatment effects but also in the plausibility of the underlying assumptions since their formal evaluation is impossible[5]. In addition, the large sample approximations on which NMA models typically rely upon fail in the presence of only a handful of studies per comparison. Hence, lack of robustness and potentially limited reliability are common issues that might be encountered when analyzing sparse treatment networks[6]. To avoid these issues, it would seem reasonable to wait until more studies become available for the outcome(s) of interest. However, the pace of trial production in these contexts is usually very slow and the chance to obtain shortly new study results is low.

The use of external evidence in the form of informative prior distributions has been suggested previously in meta-analysis, either to achieve more accurate estimation of the heterogeneity parameter[7,8] or for the incorporation of non-randomized evidence[9]. Here, we introduce a new framework that allows sharing information between two networks of treatments that pertain to different subgroups of the population. Our framework refers to the case when only aggregate data are available; in the presence of individual participant data, other approaches may be used, such as population adjustment methods[10]. Specifically, we use the results from a subgroup with a lot of direct evidence (i.e. many direct comparisons and many studies per comparison) forming a 'dense'

network to construct informative priors for the relative effects of the target subgroup forming a sparse network. This is a two-stage approach where at the first stage we synthesize the data from the dense network using a hierarchical NMA model in which we add a location parameter that shifts the distribution of the relative effects towards the sparse network's one. We then add a scale parameter that allows to further downweight the external data. At the second stage, the results from the first stage are used as prior information in the analysis of the sparse network. We inform the location parameter either through the data or using expert opinion. We illustrate our approach through an example of two subgroups of psychotic patients: a target subgroup of children-adolescents (CA) that forms a rather sparse network and a subgroup of chronic adult patients with acute exacerbation of schizophrenia, referred here as "general patients" (GP), that forms a dense network.

2. Motivating dataset

Our work is motivated by a recent article investigating the differences in the treatment effects among several subgroups of schizophrenia patients[11]. The aim of the article was to compare the subgroup of GP, defined as chronic patients with an acute exacerbation of positive symptoms, with more 'special' subgroups of patients: CA, first-episode patients, treatment-resistant patients, patients with negative symptoms, patients under substance abuse, and elderly. In contrast to the dataset for GP, all these special subgroups were informed by very limited direct evidence and therefore, the original authors, who acknowledged the problems of analyzing sparse networks, only used pairwise meta-analysis. To illustrate our approach, we use the data from two of the above subgroups: CA[12] and GP[13]. The former is considered the target subgroup which forms a sparse network and the latter is the only available subgroup with abundant direct evidence (forming a dense network) that can be safely used to elicit informative priors for CA.

The CA network comprises 19 randomized controlled trials (RCTs) published between 1973 and 2017 comparing 14 antipsychotics and placebo, and providing direct information for 21 out of 105 possible comparisons (**Figure 1 a**). Out of these 21 comparisons, two include 2 RCTs and the rest include 1 RCT only. In addition, many of these RCTs are small, with a median study sample size of 113 patients. The pace of new RCT production for this subgroup is very slow (on average one

RCT every 2.3 years), which confirms the difficulty of conducting RCTs for this group of patients. On the other hand, the network for general patients contains 255 RCTs comparing 33 antipsychotics and placebo, and forming 116 direct comparisons (**Figure 1 b**). The two networks are distinct, which means that they do not have any RCT in common. However, some drugs are included in both networks; this is a necessary condition to allow sharing information between them. The outcome of interest is the reduction of the overall schizophrenia symptoms which is measured in a continuous scale. Following the original article, we used the standardized mean difference (SMD) as effect measure since the different RCTs have used different scales.

3. Methods

3.1 Notation and general setting

Suppose we have a set of $N = n_1 + n_2$ studies. Studies $\{1,2, \dots, n_1\}$ inform the target population subgroup P_1 and form a sparse network while the rest $\{n_1 + 1, n_1 + 2, \dots, N\}$ target a population subgroup P_2 and form a dense network. Suppose also that T_1 treatments have been evaluated for population P_1 and T_2 treatments for population P_2. We denote with $T_a = T_1 \cup T_2$ the set of all available treatments and with $T_c = T_1 \cap T_2$ the set of common treatments included in both networks; the latter is the group of treatments for which information can be extrapolated from P_2 to the target subgroup P_1. Each study i provides the observed mean (change score/endpoint) y_{it_k} for the treatment $t_k \in T_a$ of arm k with $k = 1, \dots, K_i$ and K_i the total number of arms in i. For simplicity, in what follows we write the y_{it_k} as y_{ik}, its variance sd_{ik}^2, and the respective 'true' mean is denoted as θ_{ik}. We consider treatment $j = 1$ as the common network reference for both P_1 and P_2 and the relative effects μ_{1j} $(j > 1, j \in T_a)$ as the basic parameters. For every study, we arbitrarily choose the treatment of arm $k = 1$ (t_1) as the study-specific baseline treatment.

3.2 Models for sharing information between two population subgroups

Different models stem from different assumptions about the relation between the two subgroups P_1 and P_2. We start with the description of the standard hierarchical NMA model which in the present setting synthesizes the two subgroups in a naïve way, namely as if they were equivalent.

Then, we move to more plausible NMA models in which we incorporate a location parameter aiming to make P_1 and P_2 more 'similar' and a scale parameter aiming to increase the uncertainty of the studies in subgroup P_2.

3.2.1 Naïve synthesis

In this approach the two subgroups are combined together as if all participants were coming from the same population P. In other words, this approach makes the strong assumption that P_1 and P_2 are equivalent, namely $P_1 \equiv P_2 \equiv P$. Hence, for every study $i = 1, \dots N$ and treatment $t_k \in T_a$, the observed means follow a normal distribution

$$y_{ik} \sim \mathrm{N}(\theta_{ik}, sd_{ik}^2). \quad (1)$$

Then, the underlying SMD between every treatment $t_2, t_3, \dots, t_{K_i}$ versus the baseline treatment t_1 is

$$\delta_{i,1k} = \frac{\theta_{ik} - \theta_{i1}}{sd_i^{pooled}}, \quad (k > 1), \quad (2)$$

with $sd_i^{pooled} = \sqrt{\sum_{k=1}^{K_i} \frac{n_{ik} sd_{ik}^2}{n_{ik} - K_i}}$ the between-arm standard deviation. Under the random-effects assumption, the underlying relative effects are assumed to follow a multivariate normal distribution

$$\boldsymbol{\delta_i} \sim \mathrm{N}_{K_i - 1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (3)$$

where $\boldsymbol{\mu} = (\mu_{12}, \dots, \mu_{1K_i})$ is the vector of the summary relative effects and $\boldsymbol{\Sigma}$ the between-study variance-covariance matrix with entries τ^2 (the common heterogeneity variance across comparisons) in the diagonal and $\frac{\tau^2}{2}$ in the off-diagonal. Finally, the transitivity assumption for every pair of treatments $j, l \in T_a$ $(j, l > 1)$ implies that

$$\mu_{jl} = \mu_{1l} - \mu_{1j}.$$

Non-informative prior distributions are usually employed for every μ_{1j} and τ, such as $\mathrm{N}(0, 10000)$ and the half-normal $\mathrm{HN}(1)$ respectively.

Apart from the strong assumption about the similarity of the two subgroups, a further problem in the present setting is that the final results for the joint population P will be dominated by the dense network (P_2) rather than from the target subgroup P_1. The same would apply if the results of the dense network were used directly as prior information for P_1. To avoid this issue, in the proposed

approach we extrapolate the results from P_2 to P_1 before forming informative prior distributions for P_1.

3.2.2 Using the external subgroup P_2 to construct informative priors for the target subgroup P_1

This is a two-stage approach where at the first stage we extrapolate the results from the dense network of subgroup P_2 to P_1 and at the second stage we use predictions from this extrapolation to form prior distributions for the analysis of P_1.

First stage: Extrapolating external results to the target population subgroup

In contrast to the naïve synthesis, the main assumption here is that the two subgroups have some underlying differences in the treatment effects which can be considered as a different location of their outcome distributions. To incorporate some uncertainty about this assumption, in Equation (1) we introduce a scale parameter $w_i \in (0,1]$ that inflates the variance of each y_{ik} for each $i = n_1 + 1, n_1 + 2, \dots, N$. This modifies Equation (1) into

$$y_{ik} \sim \mathrm{N}\left(\theta_{ik}, \frac{sd_{ik}^2}{w_i}\right). \tag{4}$$

The location parameters $\boldsymbol{\beta}$ that shift the original distribution of the SMDs in P_2 towards the distribution of P_1 are then added in Equation (3):

$$\boldsymbol{\delta}_i \sim \mathrm{N}_{\mathrm{K_i}-1}(\boldsymbol{\mu} - \boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{5}$$

with $\boldsymbol{\beta} = \left(\beta_{12}, \dots, \beta_{1K_i}\right)$ being the vector of the 'true' differences between the comparison-specific SMDs of the two subgroups.

By fitting this modified NMA model we obtain the extrapolated SMD estimates $\hat{\mu}_{jl}^{P_2}$ for every pair of treatments $j, l \in T_c$. Subsequently, the predictive distributions of $\hat{\mu}_{jl}^{P_2^*}$, namely $\mathrm{N}\left(\hat{\mu}_{jl}^{P_2} - \hat{\beta}_{jl}, (\hat{\tau}^{P_2})^2\right)$, are used at the second stage as prior distributions for P_1.

Second stage: NMA of the target subgroup using informative prior distributions for relative effects

Here the model of Section 3.2.1 is used for $i = 1, \dots, n_1$. The key difference is that for the μ_{1l}s we use as prior distribution $\mu_{1l}^{P_1} \sim \mathrm{N}\left(\hat{\mu}_{1l}^{P_2} - \hat{\beta}_{1l}, (\hat{\tau}^{P_2})^2\right)$. The use of informative priors is expected to improve the precision in the final NMA estimates. Moreover, those informative priors are not expected to dominate the analysis as they are constructed at the first stage by extrapolating the results of the dense network to those expected for the target sparse network.

3.2.3 Informing the location and the scale parameters

Constructing prior distributions for βs using the data

To obtain an estimate of the difference in the outcome distributions between the two subgroups, we first use the estimates from pairwise meta-analysis $u_{1j}^{P_1}$ and $u_{1j}^{P_2}$. Using the difference of these estimates $d_{1j} = u_{1j}^{P_2} - u_{1j}^{P_1}$, we construct prior distributions for the parameters β_{1j}, namely

$$\beta_{1j} \sim \mathrm{N}\left(d_{1j}, var(d_{1j})\right). \quad (6)$$

For treatment comparisons not evaluated in P_2, we use non-informative priors. Here, we assume a common comparison-specific heterogeneity across the two subgroups $\left(\sigma_{1j}^{P_1}\right)^2 = \left(\sigma_{1j}^{P_2}\right)^2 = \sigma_{1j}^2$.

Constructing prior distributions for βs using expert opinion

We prepared a questionnaire and circulated it to psychiatrists with experience in treating schizophrenia for CA and GP (available in Appendix 1). We asked each expert to provide an estimate of the 'expected' mean score reduction in the PANSS scale from baseline to endpoint for each drug for CA given the pooled mean score reduction obtained from the data for GP. We further asked them to provide a measure of uncertainty around the given expected mean scores a) in the form of standard deviation and b) in a 10-point scale of their confidence. We obtained in total 22 responses which we averaged for each drug using a Bayesian meta-analysis model with weights accounting for both the standard deviation and the confidence rating. We used the inverse-variance meta-analysis model where the standard deviations provided by each expert were inflated

proportional to their confidence rating. A non-informative $\mathrm{N}(0,10000)$ and a weakly-informative HN(1) prior were used for the mean and the heterogeneity parameter respectively. To obtain the expected SMD between each treatment j and the reference ($u_{1j}^{P_1}$), we used the median of the sd_i^{pooled} across all the available studies for P_1. These were then subtracted from the respective $u_{1j}^{P_2}$ to obtain the excepted difference d_{1j}. Following Equation (6), the prior distributions for the parameters β_{1j} are $\beta_{1j} \sim \mathrm{N}\left(d_{1j}, var\left(d_{1j}\right)\right)$.

Prior distributions for the scale parameter

Dividing the variances in Equation (4) by w_i can be seen as a special case of the power prior method where a specific power is chosen for the likelihood of each study[9,14,15]. Values of w_i close to 0 reflect a serious downweight of the evidence coming from the external subgroup P_2 while values close to 1 imply a mild downweight. Typical choices of prior distributions for the scale parameter can be the beta or the uniform distribution. The parameters of these distributions should be chosen to reflect the prior beliefs regarding the specific characteristic according to which the downweight takes place. For example, a $\mathrm{Beta}(3,3)$ can be used for cases where a moderate downweight needs to apply for the external information. This distribution places its mass around 0.5 thus reflecting more uncertainty about the magnitude of our downweighting. Left-skewed beta priors (e.g. $\mathrm{Beta}(1,6)$) can be used for weight values close to 0 while on the other hand right-skewed beta priors (e.g. $\mathrm{Beta}(6,1)$) can be used for weight values close to 1. The previous downweighting schemes can be achieved with other types of prior distributions. For example, a $\mathrm{Unif}(0.4,0.6)$, a $\mathrm{Unif}(0.1,0.3)$ or a $\mathrm{Unif}(0.8,1)$ can be used for moderate, serious or mild downweight, respectively. Finally, it is generally not recommended to use priors for the scale parameter which can allow for values larger than 1. This could further increase the impact of external data and enable their dominance in the final NMA estimates.

4. Application

4.1 Implementation

We consider throughout CA being the target population subgroup P_1 and GP the external subgroup P_2. We applied 12 models employing different prior distributions for the location (β_{1j}) and scale parameters (w_i):

1. Naïve synthesis model for CA and GP with non-informative priors
2. NMA for CA with informative priors from GP
    a. data-based prior distributions for the location parameters with
        i. no downweight for any GP study ($w_i = 1$) using only treatments in T_c (No DW)
        ii. moderate downweight (i.e. $w_i \sim \text{Beta}(3,3)$) to high risk of bias GP studies using only treatments in T_c (RoB DW)
        iii. no downweight for any GP study evaluating treatments in T_c and moderate downweight (i.e. $w_i \sim \text{Beta}(3,3)$) for those evaluating at least one treatment in $T_a - T_c$ (NCT DW)
    b. prior distributions based on expert opinion for the location parameter combined the three above (i-iii) possibilities for the scale parameters
3. NMA for CA with non-informative priors (i.e. 'standard' NMA)
4. Pairwise meta-analysis for CA with non-informative priors

All the analyses were conducted using the rjags[16] package through the R statistical software (R version 4.0.3, 2020-10-10)[17]. For all models, we ran two chains in parallel, performed 50,000 iterations and discarded the first 10,000 samples of each chain. We checked the chains convergence using the Gelman-Rubin criterion[18] with a value below 1.1 to indicate failure of convergence. We also visually checked the Markov chains using trace plots to inspect the sampling behavior and assess mixing across chains and convergence.

4.2 Results

The estimates for the basic comparisons of the CA network are depicted in and for all relative comparisons in Appendix 2, Tables 9-16. As expected, for most comparisons the use of informative priors leads to a substantial improvement in the precision of the relative effects. The naïve synthesis model provides the most precise results but relies on a strong assumption that is likely to be implausible. Interestingly, CA results from the standard NMA model (with non-informative priors) tend to be less precise than the respective direct estimates. This occurs often in

sparse networks since indirect comparisons as well as heterogeneity are estimated with large uncertainty. Overall, standard NMA most often does not give any insight for comparisons without direct evidence as it yields very large credible intervals. For some of these comparisons (e.g. Haloperidol vs Placebo) the NMA models with informative priors result in more conclusive relative effect estimates. In terms of point estimates, results appear generally robust across the different models with only few exceptions, such as the extreme cases of Fluphenazine and Trifluoperazine. Those two are very old drugs for which the evidence is outdated and sparse in both GP and CA networks. Overall, point estimates appeared to be more robust for the basic comparisons for which direct evidence is available in the network of CA. Finally, no convergence issues were identified across the chains. The corresponding trace plots are available in Appendix 2.

Only small differences can be observed among the six models with informative priors suggesting that the approach of informing the location and the scale parameters does not affect materially the final predictions; this is possibly due to the huge amount of data in the GP network. For a few comparisons the data-based approach for $\boldsymbol{\beta}$ provides to some degree different relative effects than those obtained from the expert opinion approach implying that the experience from clinical practice does not fully agree with the available data. Those comparisons might need to be prioritized for the design of new trials. Of course, when the full GP network was used and only studies comparing treatments in $T_a - T_c$ were downweighed some extra precision was gained. Such an approach might be more useful when the external network is not as dense as in the present dataset.

Assessment of inconsistency was performed using the node-splitting method[5]. Inference regarding the presence of inconsistency was based on Bayesian p-values and on the visual inspection of the posterior densities of direct and indirect evidence, Bayesian p-values were calculated as $2 \times \min\{P, 1-P\}$ where P is the probability that the difference between direct and indirect comparisons is positive. Additionally, the different antipsychotics and placebo were ranked across all models using the surface under the cumulative ranking curve (SUCRA)[19] method. Overall, the respective posterior densities of the direct and indirect estimates appeared to be quite similar across the different comparisons and models and no p-value indicated lack of consistency (Appendix 2, Figures 1-5, Tables 1-8). In terms of ranking, the results appear to be generally robust across

models including the naïve synthesis and the standard NMA model. Clozapine was always ranked as the most effective antipsychotic. Finally, Olanzapine, Risperidone and Molindone were ranked in second, third and fourth position respectively across the most of the fitted models (Appendix 2, Figure 6).

5. Discussion

In this paper, we present a Bayesian framework for NMA of sparse networks aiming to improve the performance of the relative effect estimates in terms of precision and reliability. To this end, we borrow information using external data from a dense network that targets a different subgroup of the population. We model the differences between the two populations with a two-stage approach. At the first stage, we analyze only the dense network and we extrapolate its results to the sparse network though the incorporation of a location parameter in the distribution of the summary relative effects. Then, we use the extrapolated results to construct informative prior distributions for the target population subgroup. Similar approaches have been used previously at the level of primary studies[7,20–22]. We further introduce a scale parameter that inflates the variance of the studies in the dense network to reflect the potential uncertainty regarding the assumed relationship between the two population subgroups[23,24].

We proposed two different approaches for informing the location parameters; one based on the data and one based on expert opinion. The former is easier to implement but the data from the sparse network are used in both stages and this might introduce some correlation that is ignored. In addition, this approach requires that a certain number of direct comparisons is available in both networks, otherwise the use of non-informative priors would be necessary for several βs at the first stage. The prior elicitation approach is general as it does not rely on the amount of direct evidence. However, results from such an approach is always subjective to some degree. Here, we requested the necessary information from the psychiatrists though a questionnaire and some of them raised concerns that it was difficult for them to make a suggestion. Other ways to elicit information from experts exist. For example, Turner et al. used a three stage elicitation approach where the experts were interviewed through 1:1 meetings or by telephone[25].

In our motivating example, the two approaches for constructing priors for the location parameters were generally in agreement. For the scale parameters, a moderate or no downweight was chosen mostly in terms of demonstrating our approach. Compared to the standard NMA with non-informative priors, our two-stage approach provided substantially more precise estimates. The lack of robustness of the standard NMA model for this dataset is also depicted from the fact that almost always it yielded less precise results than pairwise meta-analysis. As expected, the most precise results were obtained from the naïve synthesis model. Nevertheless, this model does not account for any differences across the two population subgroups.

In the presence of few studies, estimation of the heterogeneity is challenging[8]. In our analysis we used a weakly-informative half normal prior distribution for the heterogeneity parameter[25]. Using informative priors for heterogeneity could possibly further increase the precision of our approach. Usually, these are based on empirical distributions that depend on the type of outcome and treatment comparisons[26–28]. Alternatively, more sophisticated options that allow the incorporation of external data in the estimation of the heterogeneity parameter may be applied[8].

Overall, the proposed framework offers a reliable approach for assisting the analysis of sparse networks. Of course, its application requires close collaboration between statisticians and experienced clinicians to ensure that sharing of information between different population subgroups is clinically meaningful. Finally, conducting NMA in cases of sparse networks will always be a challenging procedure and the best approach will likely be context-dependent. It is recommended that an extensive sensitivity analysis should always carried out to investigate the robustness of the findings under different analysis schemes (e.g. different choice of prior distributions or different choice of model etc.)

References

1. Chaimani A, Caldwell DM, Li T, Higgins JP, Salanti G. Undertaking network meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd; 2019:285-320. doi:10.1002/9781119536604.ch11

2. Nikolakopoulou A, Mavridis D, Furukawa TA, et al. Living network meta-analysis compared with pairwise meta-analysis in comparative effectiveness research: empirical study. *BMJ*. 2018;360. doi:10.1136/bmj.k585

3. Veroniki AA, Tsokani S, Zevgiti S, et al. Do reporting guidelines have an impact? Empirical assessment of changes in reporting before and after the PRISMA extension statement for network meta-analysis. *Syst Rev*. 2021;10(1):246. doi:10.1186/s13643-021-01780-9

4. U.S. Department of Health and Human Services Food and Drug Administration. Leveraging existing clinical data for extrapolation to pediatric uses of medical devices—Guidance for industry and Food and Drug Administration staff draft guidance. 2016.

5. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29(7-8):932-944.

6. Yoon JH, Dias S, Hahn S. A method for assessing robustness of the results of a star-shaped network meta-analysis under the unidentifiable consistency assumption. *BMC Med Res Methodol*. 2021;21(1):113. doi:10.1186/s12874-021-01290-1

7. Röver C, Wandel S, Friede T. Model averaging for robust extrapolation in evidence synthesis. *Stat Med*. 2019;38(4):674-694. doi:https://doi.org/10.1002/sim.7991

8. Turner RM, Domínguez-Islas CP, Jackson D, Rhodes KM, White IR. Incorporating external evidence on between-trial heterogeneity in network meta-analysis. *Stat Med*. 2019;38(8):1321-1335. doi:https://doi.org/10.1002/sim.8044

9. Efthimiou O, Mavridis D, Debray TPA, et al. Combining randomized and non-randomized evidence in network meta-analysis. *Stat Med*. 2017;36(8):1210-1226. doi:https://doi.org/10.1002/sim.7223

10. Phillippo DM, Dias S, Ades AE, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. *Stat Med*. 2020;39(30):4885-4911. doi:https://doi.org/10.1002/sim.8759

11. Leucht S, Chaimani A, Krause M, et al. The response of schizophrenia subgroups to different antipsychotic drugs: a systematic review and meta-analysis. *The Lancet Psych*. Published online 2022.

12. Krause M, Zhu Y, Huhn M, et al. Efficacy, acceptability, and tolerability of antipsychotics in children and adolescents with schizophrenia: A network meta-analysis. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol*. 2018;28(6):659-674. doi:10.1016/j.euroneuro.2018.03.008

13. Huhn M, Nikolakopoulou A, Schneider-Thoma J, et al. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. *Lancet Lond Engl*. 2019;394(10202):939-951. doi:10.1016/S0140-6736(19)31135-3

14. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015;34(28):3724-3749. doi:https://doi.org/10.1002/sim.6728

15. Chen MH, Ibrahim JG. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46-60. doi:10.1214/ss/1009212673

16. Plummer M. *Rjags: Bayesian Graphical Models Using MCMC. R Package Version 4-10. Https://CRAN.R-Project.Org/Package=rjags.*; 2015.

17. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2016. https://www.R-project.org/

18. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*. 1998;7(4):434-455.

19. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016

20. Dunne J, Rodriguez WJ, Murphy MD, et al. Extrapolation of adult data and other data in pediatric drug-development programs. *Pediatrics*. 2011;128(5):e1242-1249. doi:10.1542/peds.2010-3487

21. Dunoyer M. Accelerating access to treatments for rare diseases. *Nat Rev Drug Discov*. 2011;10(7):475-476. doi:10.1038/nrd3493

22. Hampson LV, Whitehead J, Eleftheriou D, Brogan P. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Stat Med*. 2014;33(24):4186-4201. doi:10.1002/sim.6225

23. Wadsworth I, Hampson LV, Jaki T. Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods. *Stat Methods Med Res*. 2018;27(2):398-413. doi:10.1177/0962280216631359

24. Hlavin G, Koenig F, Male C, Posch M, Bauer P. Evidence, eminence and extrapolation. *Stat Med*. 2016;35(13):2117-2132. doi:10.1002/sim.6865

25. Turner RM, Turkova A, Moore CL, et al. Borrowing information across patient subgroups in clinical trials, with application to a paediatric trial. *BMC Med Res Methodol.* 2022;22(1):49. doi:10.1186/s12874-022-01539-3

26. Röver C, Bender R, Dias S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Res Synth Methods.* 2021;12(4):448-474. doi:https://doi.org/10.1002/jrsm.1475

27. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol.* 2012;41(3):818-827. doi:10.1093/ije/dys041

28. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med.* 2015;34(6):984-998. doi:10.1002/sim.6381

29. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol.* 2015;68(1):52-60. doi:10.1016/j.jclinepi.2014.08.012

Acknowledgements: We are thankful to the following psychiatrists for providing expert opinion: Arango Celso, Carpenter Will, Crespo Facorro Benedicto, Davidson Michael, Dazzan Paola, Gerhard Gruender, Hasan Alkomiet, Heres Stephan, Hui Wu, Jauhar Sameer, Juckel Georg, McGorry Pat, Murray Robin, Musil Richard, Peter Natalie, Samara Myrto, Schneider-Thoma Johannes, Siafis Spyridon, Stip Emmanuel, Takeuchi Hiroyoshi, Weiser Mark, Zhu Yikang.

Figure 1: Sparse network for the population of CA (panel A) and informative network for GP (panel B). The size of the edges (lines in the networks) is proportional to the number of studies that compare the respective treatments while the size of the nodes is proportional to number of patients that received each treatment.

Figure 2: Results from the analysis of the network defined for the population of CA. The outcome of interest is the reduction of the overall schizophrenia symptoms and the effect measure is the standardized mean difference.

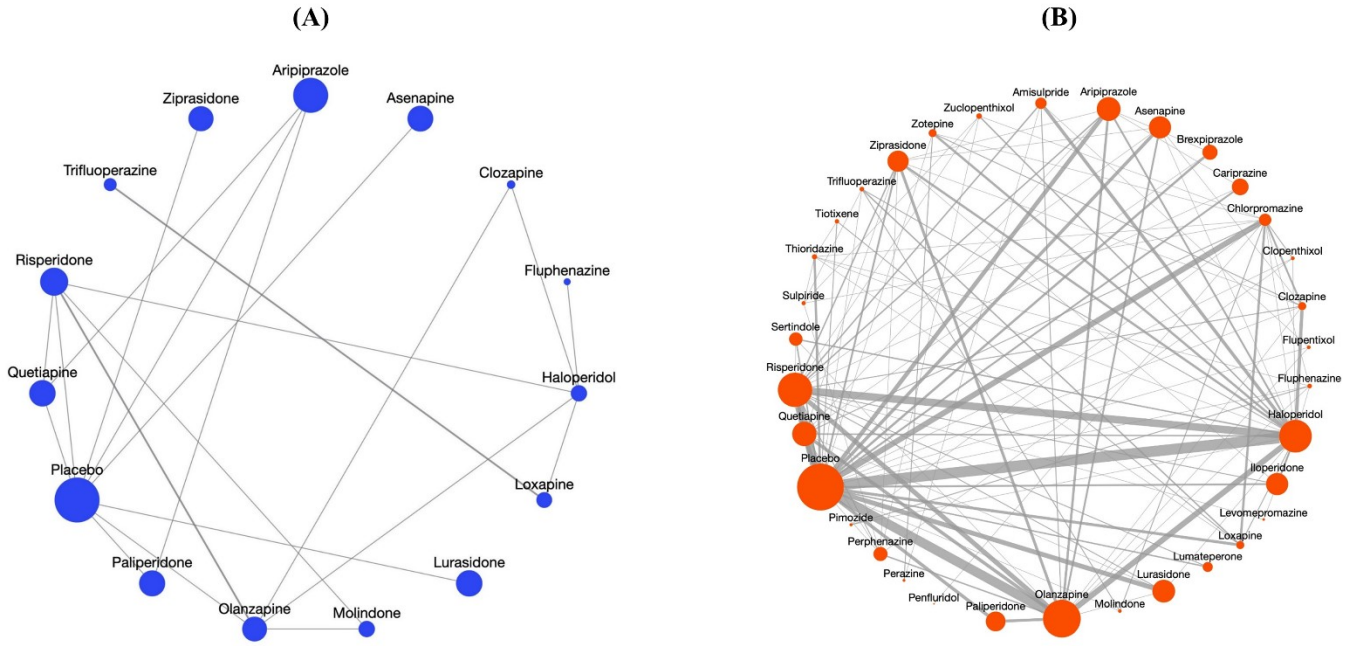

(A)
Aripiprazole
Ziprasidone
Asenapine
Trifluoperazine
Clozapine
Risperidone
Fluphenazine
Quetiapine
Haloperidol
Placebo
Loxapine
Paliperidone
Lurasidone
Olanzapine
Molindone
(B)
Amisulpride
Aripiprazole
Zuclopenthixol
Asenapine
Zotepine
Brexpiprazole
Ziprasidone
Cariprazine
Trifluoperazine
Chlorpromazine
Tiotixene
Thioridazine
Clopenthixol
Sulpiride
Clozapine
Sertindole
Flupentixol
Risperidone
Fluphenazine
Quetiapine
Haloperidol
Placebo
Iloperidone
Pimozide
Levomepromazine
Perphenazine
Loxapine
Lumateperone
Perazine
Lurasidone
Penfluridol
Olanzapine
Paliperidone
Molindone

Figure 1

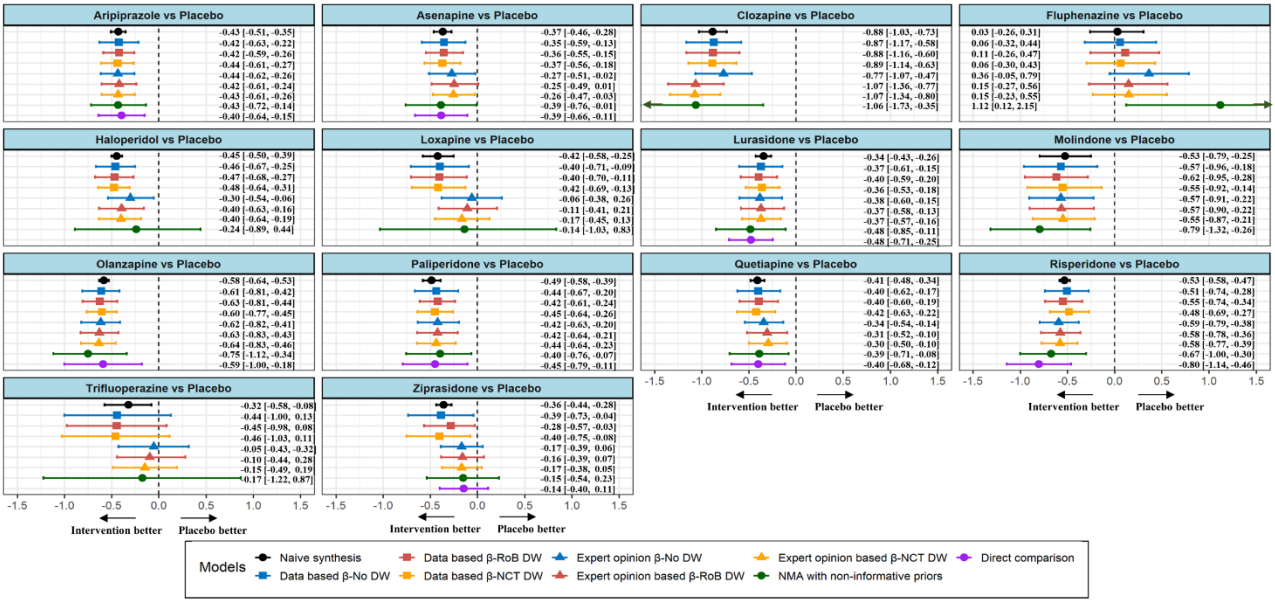

Aripiprazole vs Placebo
-0.43 [-0.51, -0.35]
-0.42 [-0.63, -0.22]
-0.42 [-0.59, -0.26]
-0.44 [-0.61, -0.27]
-0.44 [-0.62, -0.26]
-0.42 [-0.61, -0.24]
-0.43 [-0.61, -0.26]
-0.43 [-0.72, -0.14]
-0.40 [-0.64, -0.15]
Asenapine vs Placebo
-0.37 [-0.46, -0.28]
-0.35 [-0.59, -0.13]
-0.36 [-0.55, -0.15]
-0.37 [-0.56, -0.18]
-0.27 [-0.51, -0.02]
-0.25 [-0.49, 0.01]
-0.26 [-0.47, -0.03]
-0.39 [-0.76, -0.01]
-0.39 [-0.66, -0.11]
Clozapine vs Placebo
-0.88 [-1.03, -0.73]
-0.87 [-1.17, -0.58]
-0.88 [-1.16, -0.60]
-0.89 [-1.14, -0.63]
-0.77 [-1.07, -0.47]
-1.07 [-1.36, -0.77]
-1.07 [-1.34, -0.80]
-1.06 [-1.73, -0.35]
Fluphenazine vs Placebo
0.03 [-0.26, 0.31]
0.06 [-0.32, 0.44]
0.11 [-0.26, 0.47]
0.06 [-0.30, 0.43]
0.36 [-0.05, 0.79]
0.15 [-0.27, 0.56]
0.15 [-0.23, 0.55]
1.12 [0.12, 2.15]
Haloperidol vs Placebo
-0.45 [-0.50, -0.39]
-0.46 [-0.67, -0.25]
-0.47 [-0.68, -0.27]
-0.48 [-0.64, -0.31]
-0.30 [-0.54, -0.06]
-0.40 [-0.63, -0.16]
-0.40 [-0.64, -0.19]
-0.24 [-0.89, 0.44]
Loxapine vs Placebo
-0.42 [-0.58, -0.25]
-0.40 [-0.71, -0.09]
-0.40 [-0.70, -0.11]
-0.42 [-0.69, -0.13]
-0.06 [-0.38, 0.26]
-0.11 [-0.41, 0.21]
-0.17 [-0.45, 0.13]
-0.14 [-1.03, 0.83]
Lurasidone vs Placebo
-0.34 [-0.43, -0.26]
-0.37 [-0.61, -0.15]
-0.40 [-0.59, -0.20]
-0.36 [-0.53, -0.18]
-0.38 [-0.60, -0.15]
-0.37 [-0.58, -0.13]
-0.37 [-0.57, -0.16]
-0.48 [-0.85, -0.11]
-0.48 [-0.71, -0.25]
Molindone vs Placebo
-0.53 [-0.79, -0.25]
-0.57 [-0.96, -0.18]
-0.62 [-0.95, -0.28]
-0.55 [-0.92, -0.14]
-0.57 [-0.91, -0.22]
-0.57 [-0.90, -0.22]
-0.55 [-0.87, -0.21]
-0.79 [-1.32, -0.26]
Olanzapine vs Placebo
-0.58 [-0.64, -0.53]
-0.61 [-0.81, -0.42]
-0.63 [-0.81, -0.44]
-0.60 [-0.77, -0.45]
-0.62 [-0.82, -0.41]
-0.63 [-0.83, -0.43]
-0.64 [-0.83, -0.46]
-0.75 [-1.12, -0.34]
-0.59 [-1.00, -0.18]
Paliperidone vs Placebo
-0.49 [-0.58, -0.39]
-0.44 [-0.67, -0.20]
-0.42 [-0.61, -0.24]
-0.45 [-0.64, -0.26]
-0.42 [-0.63, -0.20]
-0.42 [-0.64, -0.21]
-0.44 [-0.64, -0.23]
-0.40 [-0.76, -0.07]
-0.45 [-0.79, -0.11]
Quetiapine vs Placebo
-0.41 [-0.48, -0.34]
-0.40 [-0.62, -0.17]
-0.40 [-0.60, -0.19]
-0.42 [-0.63, -0.22]
-0.34 [-0.54, -0.14]
-0.31 [-0.52, -0.10]
-0.30 [-0.50, -0.10]
-0.39 [-0.71, -0.08]
-0.40 [-0.68, -0.12]
Risperidone vs Placebo
-0.53 [-0.58, -0.47]
-0.51 [-0.74, -0.28]
-0.55 [-0.74, -0.34]
-0.48 [-0.69, -0.27]
-0.59 [-0.79, -0.38]
-0.58 [-0.78, -0.36]
-0.58 [-0.77, -0.39]
-0.67 [-1.00, -0.39]
-0.80 [-1.14, -0.46]
Trifluoperazine vs Placebo
-0.32 [-0.58, -0.08]
-0.44 [-1.00, 0.13]
-0.45 [-0.98, 0.08]
-0.46 [-1.03, 0.11]
-0.05 [-0.43, -0.32]
-0.10 [-0.44, 0.28]
-0.15 [-0.49, 0.19]
-0.17 [-1.22, 0.87]
Ziprasidone vs Placebo
-0.36 [-0.44, -0.28]
-0.39 [-0.73, -0.04]
-0.28 [-0.57, -0.03]
-0.40 [-0.75, -0.08]
-0.17 [-0.39, 0.06]
-0.16 [-0.39, 0.07]
-0.17 [-0.38, 0.05]
-0.15 [-0.54, 0.23]
-0.14 [-0.40, 0.11]
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
Intervention better
Placebo better
Models
Naive synthesis
Data based β-No DW
Data based β-RoB DW
Data based β-NCT DW
Expert opinion β-No DW
Expert opinion based β-RoB DW
Expert opinion based β-NCT DW
NMA with non-informative priors
Direct comparison

Figure 2

5. Discussion and Conclusion

In this thesis the problem of sparse data in meta-analysis was explored as this suggests a very common and complex challenge for evidence based medicine. This thesis dealt with both of the most common forms of sparse data in meta-analysis. Those refer to the case of sparse data in the form of few available studies and the case of sparse data in the form of rare events. In the next paragraphs the advantages and disadvantages of the proposed frameworks are discussed and conclusions are given based on the findings of each project.

In the first project of this thesis a new method for NMA of binary outcomes and rare events was presented. The method aimed to improve the performance of existing NMA approaches for rare events in terms of bias and precision by using the penalized likelihood function that was originally proposed by Firth for individual studies. This approach can only provide odds ratios. However, in the context of rare events, the differences between odds ratios and risk ratios are often negligible. The proposed approach was evaluated and compared with other methods through both an extended simulation study and through applications into two real life clinical examples. The proposed PL-NMA model appeared to have overall the best performance in terms of bias, especially in situations with very few studies per comparison and very low control group risks. On the other hand, the IV-NMA method was found to be a suboptimal choice of model as its results were suffering from a high amount of bias in most scenarios. The MH-NMA method was found to perform well in terms of bias and being a reliable method for analyzing rare events. In terms of Bayesian methods, it was found that the choice of weakly informative priors for the heterogeneity parameter can materially affect and improve the performance of a Bayesian logistic regression model for the analysis of rare events. Both through the simulation and the analysis of the real life

clinical example it was depicted that in presence of rare events even a choice of a non-informative prior can seriously affect the final NMA estimates. This validates a previously raised concern suggesting that Bayesian models should be fitted with cautiousness in cases of rare events.

An additional property of the proposed PL-NMA model is the ability to synthesize all studies within a network of interventions even when zero events are reported in one or all study arms. The latter can potentially be a very important property since excluding such studies from the analysis may result in sparse or disconnected networks. To date, the optimal way to treat such studies in pairwise meta-analysis or NMA is still unclear with some researchers arguing that they are informative[48] and others that those studies are non-informative and problematic[14,94]. This question was addressed through the final scenario of the simulation study. The main finding regarding the proposed PL-NMA framework suggests that either the inclusion or the exclusion of studies reporting zero events in all their arms is not having an important contribution in the final NMA estimates in terms of ORs. However, the choice of including those studies was found to increase the performance of Wald-type confidence intervals in terms of coverage probability.

Future work regarding the problem of rare events could focus on extending the literature of the proposed methodologies from pairwise meta-analysis (e.g. the beta-binomial model[48]) into the framework of NMA and compare them through simulation studies. Moreover, the issue of properly estimating the heterogeneity parameter in presence of rare events is still an open field of research[94]. In the first project of this thesis a heterogeneity parameter was added into the PL-NMA method using a two-stage approach where heterogeneity is treated as an overdispersion parameter. This framework was investigated and compared with other random-effects NMA models. The findings suggested that none of the fitted models was performing considerably well for estimating heterogeneity thus more work is needed for addressing this issue in presence of rare events. Finally,

more extended simulation studies are needed to address the questions regarding the inclusion or exclusion of studies which report zero events in one or all study arms as the simulation conducted in the first project addressed this question only through one scenario.

The second project of this thesis allowed for the broad application of the methods suitable for meta-analysis of rare events in the context of Covid-19 trials. Specifically, based on the database of the COVID-NMA initiative the metaCOVID R-Shiny application was constructed to allow all the end-users of COVID-NMA platform and other external users to easily conduct their own meta-analyses of Covid-19 trials based on the most up-to-date data. The issue of rare events appears to be a common problem for Covid-19 trials as many of the binary primary outcomes investigated through COVID-NMA are rarely observed. Results posted in the COVID-NMA platform are obtained based only on the IV model which can be problematic for analyzing rare events. However, the replacement of this IV model and the presentation of sensitivity analysis results can be very challenging as this could make the COVID-NMA platform hard to be used. On the other hand, through metaCOVID the users can easily fit additional models such as the PL-NMA model or the MH model which are more suitable for analyzing rare events. The latter can take place through a user-friendly environment which does not require any programming expertise by the users.

The functionality of the application goes beyond the analysis of rare events as it allows for a meta-analysis to be performed based on several analysis criteria such as the method of estimating the heterogeneity, the type of meta-analytical model (i.e. common or random effect(s)) or the set of trials which are included in the meta-analysis (e.g. exclusion or not of trials based on their risk of bias or publication status). Moreover, several subgroup analysis criteria are available in this tool such as subgroup analysis according to the reported severity of the patients in the trials or based on the trial funding status.

All living evidence synthesis projects are highly resource- and time-demanding as they involve large amounts of data that need to be continuously updated and analyzed. Rapid and efficient automation tools like metaCOVID are required for the sustainability of any living systematic review that aims to have a very short delay (i.e. a few weeks) between every update as well as to allow for immediate dissemination and access to the findings. Finally, although metaCOVID is linked with the COVID-NMA initiative the application can be seen as a pilot version which based on future work it could become a generalized application that could be tailored for living meta-analyses of any condition.

The third project of this thesis addressed the issue of sparse networks of interventions. In this project a Bayesian approach was proposed to improve the performance of the NMA estimates in terms of precision and reliability. To do so, a two-stage approach was proposed. This aimed to analyze a sparse network while borrowing strength from a relevant dense network of interventions which pertains into a different subgroup of the population. A modified NMA model is introduced at the first-stage of the approach. This model analyzes the dense network of interventions with the main purpose to properly extrapolate the results of this network to those expected for the sparse network. The modifications refer to a location and a scale parameter which are added to the standard NMA model. The location parameter models the differences between the relative effects of the two populations while the scale parameter inflates the variances of the studies in the dense network. The predictive distributions of the extrapolated NMA estimates obtained at the first-stage are used as informative priors for the analysis of the sparse network at the second-stage. This analysis takes place using those informative priors for the basic parameters of the model. The latter is the key-difference between the two-stage approach and the standard NMA model which places non-informative priors for those basic parameters.

The new framework was illustrated through the analysis of a sparse network of psychiatric patients which refers to the population of the children and adolescents. The dense network used for the illustrative example refers to the population of general patients where substantially more evidence is available. A prior distribution for the added location parameter was constructed based both on a data-based approach and an elicitation of expert opinion. The findings of this analysis suggest that the two-stage framework can materially improve the performance of the NMA estimates in terms of precision and robustness as this model allows for using external evidence and also adjusts for the inevitable differences that exist between the two populations which participate in the process of sharing information.

The proposed two-stage framework appeared to facilitate the estimation of intervention effects in cases of sparse networks. However, the method has its own limitations. The proposed framework relies on the existence of a dense network which can be used as external evidence for the analysis of the sparse network. The latter is not always possible thus in that case the only evidence which is available arises from a small set of studies. In such circumstances the proper way for analyzing sparse networks should be further investigated both in the frequentist and the Bayesian framework of NMA. Moreover, in cases where potential dense networks exist then the choice of sharing information or not should always depend on expert or clinical opinion about the similarities or not between the two populations. For example, it is suggested the two-stage process should not be used in cases of substantial clinical differences between the two populations. Finally, the two-stage process focused mostly on the proper estimation of the summary intervention effects. However, the estimation of the heterogeneity parameter can also be challenging in presence of only few available studies. In this direction the solutions of using empirical or weakly informative priors for the heterogeneity parameter in NMA could be further investigated as a future work.

In conclusion, the proposed approaches for dealing with sparse data either in the form of rare events or in the form of few available studies provide reliable alternatives to the existing limited literature for NMA. However, the analysis of sparse data in meta-analysis is always a very challenging procedure and thus the reliability of the final summary estimates should always be investigated through extensive sensitivity analysis. The latter implies that in such cases no method should be thought as the uniquely best approach and several analysis schemes should always be considered in order to investigate the robustness of the final estimates.

References

1. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *bmj*. 2016;352:i1981.

2. Greenland S, Schwartzbaum J, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *American journal of epidemiology*. 2000;151 5:531-539.

3. Turner RM, Turkova A, Moore CL, et al. Borrowing information across patient subgroups in clinical trials, with application to a paediatric trial. *BMC Medical Research Methodology*. 2022;22(1):49. doi:10.1186/s12874-022-01539-3

4. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet*. 2005;365(9465):1159-1162. doi:10.1016/S0140-6736(05)71879-1

5. Vandermeer B, Bialy L, Hooton N, et al. Meta-analyses of safety data: a comparison of exact versus asymptotic methods. *Stat Methods Med Res*. 2009;18(4):421-432. doi:10.1177/0962280208092559

6. Nelder JA, Wedderburn RWM. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*. 1972;135(3):370-384. doi:10.2307/2344614

7. Agresti A. *Categorical Data Analysis*. John Wiley & Sons; 2003.

8. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*. 2009;9(1):56.

9. Cox DR, Snell EJ. A General Definition of Residuals. *Journal of the royal statistical society series b-methodological*. 1968;30:248-265.

10. King G, Zeng L. Logistic Regression in Rare Events Data. *Political Analysis*. 2001;9(2):137-163. doi:10.1093/oxfordjournals.pan.a004868

11. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1-10.

12. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. *American journal of epidemiology*. 2018;187(4):864-870.

13. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA*. 1995;274(22):1800-1804. doi:10.1001/jama.274.22.1800

14. Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta-analyses. *Cochrane handbook for systematic reviews of interventions*. Published online 2019:241-284.

15. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008;17(3):279-301. doi:10.1177/0962280207080643

16. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*. 2002;21(11):1539-1558.

17. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*. 2010;1(2):97-111.

18. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Bmj*. 2003;327(7414):557-560.

19. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons; 2011.

20. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med*. 1996;15(6):619-629. doi:10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A

21. 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. *Medical Decision Making*. 2013;33(5):607-617.

22. Raudenbush SW. Analyzing effect sizes: Random-effects models. In: *The Handbook of Research Synthesis and Meta-Analysis, 2nd Ed.* Russell Sage Foundation; 2009:295-315.

23. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials*. 1986;7(3):177-188.

24. Paule RC, Mandel J. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*. 1982;87(5):377-385.

25. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*. 2005;30(3):261-293. doi:10.3102/10769986030003261

26. Röver C. Bayesian Random-Effects Meta-Analysis Using the bayesmeta R Package. *Journal of Statistical Software*. 2020;93(6):1-51. doi:10.18637/jss.v093.i06

27. Borenstein M, Higgins JP. Meta-analysis and subgroups. *Prevention science*. 2013;14(2):134-143.

28. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med*. 2004;23(11):1663-1682. doi:10.1002/sim.1752

29. Seide SE, Jensen K, Kieser M. A comparison of Bayesian and frequentist methods in random-effects network meta-analysis of binary data. *Research synthesis methods*. 2020;11(3):363-378.

30. Turner RM, Bird SM, Higgins JP. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PloS one*. 2013;8(3):e59202.

31. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods*. 2017;8(1):79-91. doi:10.1002/jrsm.1217

32. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? *Biom J*. 2018;60(6):1040-1058. doi:10.1002/bimj.201800071

33. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984-998. doi:10.1002/sim.6381

34. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*. 2016;7(1):55-79. doi:10.1002/jrsm.1164

35. Petropoulou M, Mavridis D. A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study. *Statistics in medicine*. 2017;36(27):4266-4280.

36. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*. 2010;36(3):48.

37. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health*. 2019;22(4):153-160. doi:10.1136/ebmental-2019-300117

38. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemp Clin Trials*. 2014;37(1):129-138. doi:10.1016/j.cct.2013.11.012

39. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in medicine*. 2007;26(9):1964-1981.

40. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2005;54(2):367-384. doi:https://doi.org/10.1111/j.1467-9876.2005.00489.x

41. Efthimiou O. Practical guide to the meta-analysis of rare events. *Evid Based Ment Health*. 2018;21(2):72-76. doi:10.1136/eb-2018-102911

42. J. Sweeting M, J. Sutton A, C. Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine*. 2004;23(9):1351-1375.

43. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*. 2007;26(1):53-77.

44. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. *Stat Med*. 2010;29(12):1259-1265. doi:10.1002/sim.3607

45. Schmid CH, Stijnen T, White I. *Handbook of Meta-Analysis*. CRC Press; 2020.

46. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059-1085. doi:10.1002/sim.7588

47. Keus F, Wetterslev J, Gluud C, Gooszen HG, van Laarhoven CJHM. Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. *Am J Gastroenterol*. 2009;104(3):546-551. doi:10.1038/ajg.2008.22

48. Kuss O. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. *Statistics in Medicine*. 2015;34(7):1097-1116. doi:https://doi.org/10.1002/sim.6383

49. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*. 1959;22(4):719-748.

50. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases*. 1985;27(5):335-371.

51. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. *Statistical methods in medical research*. 2016;25(6):2858-2877.

52. Senn S. Hans van Houwelingen and the Art of Summing up. *Biometrical Journal*. 2010;52(1):85-94. doi:https://doi.org/10.1002/bimj.200900074

53. Günhan BK, Röver C, Friede T. Random-effects meta-analysis of few studies involving rare events. *Research Synthesis Methods*. 2020;11(1):74-90. doi:10.1002/jrsm.1370

54. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in medicine*. 2010;29(29):3046-3067.

55. Cai T, Parast L, Ryan L. Meta-Analysis For Rare Events. *Stat Med*. 2010;29(20):2078-2089. doi:10.1002/sim.3964

56. Bhaumik DK, Amatya A, Normand SL, et al. Meta-Analysis of Rare Binary Adverse Event Data. *J Am Stat Assoc*. 2012;107(498):555-567. doi:10.1080/01621459.2012.664484

57. Chaimani A, Caldwell DM, Li T, Higgins JP, Salanti G. Undertaking network meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd; 2019:285-320. doi:10.1002/9781119536604.ch11

58. Higgins JP, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? *The Lancet*. 2015;386(9994):628-630.

59. Efthimiou O, Debray TP, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. *Research synthesis methods*. 2016;7(3):236-263.

60. Chaimani A, Higgins JP, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PloS one*. 2013;8(10):e76654.

61. Afach S, Chaimani A, Evrenoglou T, et al. Meta-analysis results do not reflect the real safety of biologics in psoriasis. *British Journal of Dermatology*. Published online 2020.

62. 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(6):683-691. doi:10.1016/s0895-4356(97)00049-8

63. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research synthesis methods*. 2012;3(2):80-97.

64. Donegan S, Williamson P, Gamble C, Tudur-Smith C. Indirect Comparisons: A Review of Reporting and Methodological Quality. *PLOS ONE*. 2010;5(11):e11054. doi:10.1371/journal.pone.0011054

65. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*. 2006;101(474):447-459.

66. 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 Medicine*. 2013;11(1):159. doi:10.1186/1741-7015-11-159

67. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in medicine*. 2010;29(7-8):932-944.

68. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med*. 2002;21(16):2313-2324. doi:10.1002/sim.1201

69. HIGGINS JPT, WHITEHEAD A. BORROWING STRENGTH FROM EXTERNAL TRIALS IN A META-ANALYSIS. *Statistics in Medicine*. 1996;15(24):2733-2749. doi:https://doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0

70. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*. 2009;10(4):792-805. doi:10.1093/biostatistics/kxp032

71. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *Bmj*. 2005;331(7521):897-900.

72. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Synth Methods*. 2012;3(2):111-125. doi:10.1002/jrsm.1045

73. Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Stat Methods Med Res*. 2013;22(2):133-158. doi:10.1177/0962280211432219

74. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med*. 2011;30(20):2481-2498. doi:10.1002/sim.4172

75. Rücker G. Network meta-analysis, electrical networks and graph theory. *Res Synth Methods*. 2012;3(4):312-324. doi:10.1002/jrsm.1058

76. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003;326(7387):472. doi:10.1136/bmj.326.7387.472

77. Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *J Clin Epidemiol*. 2010;63(8):875-882. doi:10.1016/j.jclinepi.2009.08.025

78. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Med Decis Making*. 2013;33(5):641-656. doi:10.1177/0272989X12455847

79. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Medical Research Methodology*. 2013;13(1):35. doi:10.1186/1471-2288-13-35

80. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods*. 2012;3(2):98-110. doi:10.1002/jrsm.1044

81. Jackson D, Barrett JK, Rice S, White IR, Higgins JPT. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat Med*. 2014;33(21):3639-3654. doi:10.1002/sim.6188

82. White IR. Network Meta-analysis. *The Stata Journal*. 2015;15(4):951-985. doi:10.1177/1536867X1501500403

83. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016

84. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC medical research methodology*. 2015;15(1):58.

85. Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clin Epidemiol*. 2014;6:451-460. doi:10.2147/CLEP.S69660

86. Soares MO, Dumville JC, Ades AE, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2014;177(1):259-279. doi:10.1111/rssa.12010

87. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Stat Med*. 2013;32(17):2935-2949. doi:10.1002/sim.5764

88. Turner RM, Domínguez-Islas CP, Jackson D, Rhodes KM, White IR. Incorporating external evidence on between-trial heterogeneity in network meta-analysis. *Statistics in Medicine*. 2019;38(8):1321-1335. doi:https://doi.org/10.1002/sim.8044

89. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol*. 2012;41(3):818-827. doi:10.1093/ije/dys041

90. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68(1):52-60. doi:10.1016/j.jclinepi.2014.08.012

91. Röver C, Wandel S, Friede T. Model averaging for robust extrapolation in evidence synthesis. *Statistics in Medicine*. 2019;38(4):674-694. doi:https://doi.org/10.1002/sim.7991

92. Yoon JH, Dias S, Hahn S. A method for assessing robustness of the results of a star-shaped network meta-analysis under the unidentifiable consistency assumption. *BMC Medical Research Methodology*. 2021;21(1):113. doi:10.1186/s12874-021-01290-1

93. 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(1):e86754-e86754. doi:10.1371/journal.pone.0086754

94. Efthimiou O, Rücker G, Schwarzer G, Higgins JP, Egger M, Salanti G. Network meta-analysis of rare events using the Mantel-Haenszel method. *Statistics in medicine*. 2019;38(16):2992-3012.

95. Schwarzer G, Efthimiou O, Rücker G. Inconsistent results for Peto odds ratios in multi-arm studies, network meta-analysis and indirect comparisons. *Research Synthesis Methods*. 2021;12(6):849-854. doi:10.1002/jrsm.1503

96. QUENOUILLE MH. NOTES ON BIAS IN ESTIMATION. *Biometrika*. 1956;43(3-4):353-360. doi:10.1093/biomet/43.3-4.353

97. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM; 1982.

98. Kosmidis I. Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics*. 2014;6(3):185-196. doi:10.1002/wics.1296

99. Cordeiro GM, McCullagh P. Bias Correction in Generalized Linear Models. *Journal of the Royal Statistical Society Series B (Methodological)*. 1991;53(3):629-643.

100. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. Published online 1993:27-38.

101. Jeffreys H. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences*. 1946;186(1007):453-461.

102. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in medicine*. 2002;21(16):2409-2419.

103. Kosmidis I, Firth D. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*. Published online 2020. https://doi.org/10.1093/biomet/asaa052

104. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine*. 1999;18(20):2693-2708.

105. Kulinskaya E, Olkin I. An overdispersion model in meta-analysis. *Statistical Modelling*. 2014;14(1):49-76.

106. Boutron I, Chaimani A, Meerpohl JJ, et al. The COVID-NMA Project: Building an Evidence Ecosystem for the COVID-19 Pandemic. *Ann Intern Med*. 2020;173(12):1015-1017. doi:10.7326/M20-5261

107. Boutron I, Chaimani A, Devane D, et al. Interventions for the treatment of COVID-19: a living network meta-analysis. *Cochrane Database of Systematic Reviews*. 2020;(11). doi:10.1002/14651858.CD013770

108. Boutron I, Chaimani A, Devane D, et al. Interventions for the prevention and treatment of COVID-19: a living mapping of research and living network meta-analysis. *Cochrane Database of Systematic Reviews*. 2020;(11). doi:10.1002/14651858.CD013769

109. Siemieniuk RA, Bartoszko JJ, Ge L, et al. Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ*. 2020;370:m2980.

110. Juul S, Nielsen N, Bentzer P, et al. Interventions for treatment of COVID-19: a protocol for a living systematic review with network meta-analysis including individual patient data (The LIVING Project). *Syst Rev*. 2020;9(1):108. doi:10.1186/s13643-020-01371-0

111. Morris CN. Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*. 1983;78(381):47-55. doi:10.1080/01621459.1983.10477920

Table of contents

Annexes

List of annexes

Annex 1: Supplementary files for article in Part 1

Data generation mechanism for heterogeneous datasets

For generating heterogeneous datasets, we need to generate the study-specific true $logOR$s ($\delta_{i,1k}$) for the comparison of the $k \in K_i$ treatment versus the baseline treatment 1 in study i. To do so, we use the multivariate normal distribution, hence $\delta_{i,1k} \sim MVN(d_{1k}, \mathbf{\Sigma_i})$, $k \in K_i$. The matrix $\mathbf{\Sigma_i}$ is the variance-covariance matrix of the random-effects with dimension $T_i \times T_i$, where T_i represents the number of treatments in study i. The entries of the matrix $\mathbf{\Sigma_i}$ are assumed to be equal to τ^2 for its diagonal elements and $\frac{\tau^2}{2}$ for its off-diagonal elements. For example, for a 3-arm study after following the general procedure the true $logOR$s will have values equal to 0, 0.5 and 1. Then assuming than $\tau = 0.1$, the $\delta_{i,1k}$ will be drawn from the multivariate normal distribution with mean equal to $(0, 0.5, 1)'$ and variance-covariance matrix $\mathbf{\Sigma_i}$ equal to $\begin{pmatrix} 0.01 & 0.005 & 0.005 \\ 0.005 & 0.01 & 0.005 \\ 0.005 & 0.005 & 0.01 \end{pmatrix}$.

Afterwards we can easily generate the odds, event probabilities and finally the number of events per study arm following the procedure described in the main manuscript.

Evaluated models

We evaluated the following models: the common and random-effect(s) inverse-variance (IV) model, the Mantel-Haenszel (MH) and non-central hypergeometric (NCH) common-effect models, our penalized likelihood NMA model (PL-NMA), the Binomial-Normal (BN-NMA) model and finally the two Bayesian common and random-effect(s) logistic regression models. A detailed description of our PL-NMA model exists in the main manuscript. In the main manuscript

a description exists as well for the BN-NMA model and the Bayesian models. Here, we provide a brief description for the rest of the models.

Inverse-variance model

Let $\boldsymbol{y}$ denote the vector of the observed $logOR$s in the N studies that form a network with T treatments. Then, this model can be summarized by the equation

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon} + \boldsymbol{\zeta}$$

where $\mathbf{X}$ is the design matrix with entries 0, 1, and -1 that depends on the structure of the network, $\boldsymbol{\delta}$ is the vector of the true $logOR$s representing the relative effects of all the treatments in the network versus the reference treatment, $\boldsymbol{\varepsilon}$ is the vector of the random errors, and $\boldsymbol{\zeta}$ is the vector of the random-effects. The random errors and the random-effects are assumed to follow multivariate normal distributions $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \mathbf{S})$ and $\boldsymbol{\zeta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{S}$ is a block-diagonal matrix representing the within-study variance-covariance matrix and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the random-effects. Note that the within-study variances in $\mathbf{S}$ are treated as fixed and known. The NMA estimates and their variance are obtained by, $\widehat{\boldsymbol{\delta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{y}$ and $\mathrm{V}(\widehat{\boldsymbol{\delta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ respectively, where $\mathbf{W} = (\mathbf{S} + \boldsymbol{\Sigma})^{-\mathbf{1}}$. By setting the matrix $\boldsymbol{\Sigma}$ to be the zero matrix the above random-effects model becomes the common-effect IV model.

Mantel-Haenszel model

The model can be described in 3 stages. At the first-stage, all studies are grouped according to their treatment design. At the second-stage, the following procedure takes place for each of the different treatment designs. Within the design q with N_q studies and T_q treatments, the model uses the generalized expressions for calculating the MH odds-ratios and provides the relative

treatment effect of the treatment B versus the design reference treatment A in design q ($\hat{\theta}_{q,AB}$). That is,

$$\hat{\theta}_{q,\mathrm{AB}} = \frac{L_{A+} - L_{B+}}{T_q}$$

where $L_{\mathrm{A+}} = \sum_{J=1}^{Tq} L_{AJ}$, $L_{AB}=\ln\left(\frac{C_{\mathrm{AB}}}{C_{BA}}\right)$, $C_{AB}=\sum_{i=1}^{N_q} l_{ABi}$, and $l_{ABi} = \frac{r_{Ai}(n_{Bi}-r_{Bi})}{n_{+i}}$, with r_{Ai} being the number of events for treatment A in study i and design q, n_{Bi} the sample size of treatment B in study i and design q, and n_{+i} the total sample size for study i of design q. For more information on how we can calculate the variance-covariance within each design ($\mathbf{V_q}$) we refer to the original publication. At the end of stage two, we have obtained the vector $\widehat{\boldsymbol{\theta}}$ which contains the $\widehat{\boldsymbol{\theta}}_{\boldsymbol{q}}$ for all designs q and the variance-covariance matrix $\mathbf{V}$ which is a block-diagonal matrix with diagonal elements the different sub-matrices $\mathbf{V_q}$ and off-diagonal elements being equal to zero. At the final stage, we calculate the final NMA estimates ($\widehat{\boldsymbol{\delta}}_{MH}$) by synthesizing the MH odds-ratios across all the designs using, that $\widehat{\boldsymbol{\delta}}_{MH} = (\mathbf{X'VX})^{-1}\mathbf{X'V}\boldsymbol{\theta}$ with variance-covariance matrix $\mathbf{V}\left(\widehat{\boldsymbol{\delta}}_{MH}\right) = (\mathbf{X'VX})^{-1}$.

Non-central hypergeometric model

This model is based on the within-study distributional assumption of non-central hypergeometric distribution. Specifically, given the total number of events R_i in a study i, i.e. $R_i = \sum_{k\in K_i} r_{ik}$, the vector $\boldsymbol{r} = [r_{ik}], k \in K_i - \{1\}$ follows a multivariate non-central hypergeometric distribution. For a 3-arm study the likelihood function can be written as

$$\frac{\exp(r_{i2}\delta_{12} + r_{i3}\delta_{13})}{\sum_{j^{(2)},j^{(3)}} \binom{n_{i2}}{j^{(2)}}\binom{n_{i3}}{j^{(3)}}\binom{n_{i1}}{R_i - j^{(2)} - j^{(3)}} \exp(j^{(2)}\delta_{12} + j^{(3)}\delta_{13})}$$

where the sum in the denominator is over all pairs $(j^{(2)}, j^{(3)})$ for which the binomial coefficients can be defined. The version implemented in netmeta and used in our simulations is based on the Breslow's approximation to the above likelihood which is valid when the outcome of interest is rare. Using this approximation, the likelihood function becomes

$$\frac{\exp(r_{i2}\delta_{12} + r_{i3}\delta_{13})}{[(n_{i1} + n_{i2}\exp(\delta_{12}) + n_{i3}\exp(\delta_{13})]^{R_i}}$$

S-Table 1: Results in terms of mean bias, mean square error (MSE) and coverage probability (%) across all 32 scenarios and across all frequentist models

#	IV-Common			IV-Random			MH			NCH			PL-NMA-Common			PL-NMA- Random			Binomial Normal			Logistic NMA (Standard likelihood-Common)		
	Mean bias	MSE	Coverage (%)	Mean bias	MSE	Coverage (%)	Mean bias	MSE	Coverage (%)	Mean bias	MSE	Coverage (%)	Mean bias	MSE	Coverage (Wald- **Prof. Lik, %)**	Mean bias	MSE	Coverage (%)	Mean bias	MSE	Coverage (%)	Mean bias	MSE	Coverage (%)
1	-0.07	0.34	97.80	-0.07	0.35	98.00	-0.04	0.37	96.90	-0.04	0.34	96.20	0.01	0.35	95.00 **95.30**	0.01	0.35	95.20	0.04	0.38	95.00	0.04	0.38	95.00
2	-0.07	0.35	97.30	-0.07	0.36	97.30	-0.05	0.37	96.40	-0.05	0.35	95.60	0.00	0.36	94.00 **94.50**	0.00	0.36	94.00	0.03	0.40	94.40	0.03	0.40	94.40
3	-0.03	0.20	97.00	-0.03	0.21	97.30	0.00	0.20	95.80	-0.05	0.18	95.40	0.01	0.20	94.60 **94.70**	0.01	0.20	95.00	0.03	0.21	94.50	0.03	0.21	94.50
4	-0.04	0.21	96.40	-0.04	0.22	96.60	-0.01	0.21	95.40	-0.06	0.18	94.60	0.00	0.20	94.20 **94.30**	0.00	0.21	94.60	0.02	0.21	93.80	0.02	0.21	93.80
5	-0.09	0.17	97.40	-0.08	0.17	97.40	-0.01	0.17	95.70	-0.03	0.16	95.40	0.00	0.17	94.50 **95.00**	0.00	0.17	94.50	0.02	0.18	94.90	0.02	0.18	94.90
6	-0.07	0.17	97.30	-0.07	0.18	97.40	0.00	0.18	95.60	-0.02	0.17	95.10	0.01	0.17	94.20 **94.80**	0.01	0.17	94.20	0.03	0.19	94.50	0.03	0.19	94.50
7	-0.05	0.10	96.50	-0.05	0.10	96.80	0.00	0.10	95.40	-0.06	0.09	94.40	0.00	0.10	94.80 **95.20**	0.00	0.10	95.00	0.01	0.10	94.70	0.01	0.10	94.70
8	-0.04	0.10	96.20	-0.04	0.10	96.50	0.00	0.10	95.00	-0.06	0.09	94.30	0.00	0.10	94.30 **94.60**	0.00	0.10	94.30	0.02	0.10	94.30	0.02	0.10	94.30
9	-0.08	0.22	97.80	-0.08	0.22	97.80	-0.06	0.22	96.90	-0.06	0.21	96.00	-0.01	0.21	94.80 **95.50**	-0.01	0.21	94.80	0.01	0.23	95.30	0.01	0.23	95.30
10	-0.07	0.21	98.00	-0.07	0.22	98.10	-0.04	0.22	97.20	-0.04	0.20	96.00	0.01	0.21	94.50 **95.30**	0.01	0.21	94.50	0.03	0.23	95.00	0.03	0.23	95.00
11	-0.07	0.27	97.60	-0.06	0.27	97.70	-0.03	0.28	96.70	-0.01	0.27	95.40	0.00	0.26	94.70 **95.00**	0.00	0.27	94.80	0.02	0.29	94.90	0.02	0.29	94.90
12	-0.06	0.27	97.10	-0.06	0.28	97.20	-0.03	0.28	96.30	-0.01	0.27	95.20	0.00	0.27	94.10 **94.60**	0.00	0.27	94.40	0.02	0.29	94.60	0.02	0.29	94.60
13	-0.12	0.48	98.30	-0.12	0.48	98.30	-0.12	0.61	97.70	-0.10	0.59	97.00	0.00	0.52	94.50 **95.00**	0.00	0.52	94.50	0.04	0.62	95.20	0.04	0.62	95.20
14	-0.12	0.48	98.20	-0.12	0.48	98.20	-0.12	0.61	97.40	-0.09	0.60	96.40	0.00	0.53	94.00 **94.80**	0.00	0.53	94.00	0.04	0.63	94.90	0.04	0.63	94.90
15	-0.14	0.25	97.90	-0.14	0.25	97.90	-0.05	0.27	96.50	-0.02	0.27	95.40	-0.02	0.24	94.00 **95.20**	-0.02	0.24	94.00	0.00	0.29	95.10	0.00	0.29	95.10
16	-0.12	0.24	97.80	-0.12	0.24	97.80	-0.02	0.27	96.60	0.01	0.27	95.10	0.01	0.25	93.30 **94.50**	0.01	0.25	93.30	0.03	0.28	94.40	0.03	0.28	94.40
17	-0.07	0.17	96.40	-0.07	0.18	96.80	0.04	0.19	95.60	0.01	0.18	95.60	0.02	0.18	95.40 **95.60**	0.02	0.18	95.60	-	-	-	0.03	0.19	95.60
18	-0.09	0.17	96.20	-0.09	0.17	96.60	0.01	0.18	95.40	-0.01	0.17	95.60	-0.01	0.17	95.20 **95.20**	-0.01	0.17	95.30	-	-	-	0.01	0.18	95.40
19	-0.14	0.30	97.20	-0.14	0.30	97.20	0.05	0.40	96.40	0.05	0.39	96.40	0.02	0.34	96.20 **95.80**	0.02	0.34	96.20	-	-	-	0.06	0.40	96.30
20	-0.15	0.30	97.50	-0.15	0.30	97.60	0.03	0.37	96.00	0.02	0.37	96.00	-0.01	0.33	95.60 **95.50**	-0.01	0.33	95.60	-	-	-	0.03	0.37	96.00
21	-0.04	0.10	96.20	-0.03	0.10	96.60	0.02	0.10	95.80	-0.02	0.10	95.60	0.01	0.10	95.80 **96.00**	0.01	0.10	96.20	-	-	-	0.02	0.10	95.80
22	-0.05	0.10	96.10	-0.05	0.10	96.60	0.00	0.10	95.50	-0.04	0.09	95.40	0.00	0.10	95.40 **95.60**	0.00	0.10	95.70	-	-	-	0.00	0.10	95.40
23	-0.01	0.06	95.30	-0.01	0.06	96.30	0.01	0.06	94.80	-0.06	0.05	92.10	0.01	0.06	94.80 **94.90**	0.01	0.06	95.50	-	-	-	0.02	0.06	94.80
24	-0.02	0.05	95.70	-0.02	0.06	96.60	0.00	0.05	95.30	-0.07	0.05	92.40	0.00	0.05	95.20 **95.40**	0.00	0.06	95.50	-	-	-	0.00	0.05	95.30
25	-0.07	0.18	95.60	-0.07	0.18	95.90	0.02	0.19	95.20	0.00	0.19	95.40	0.00	0.18	95.30 **95.20**	0.00	0.18	95.30	-	-	-	0.02	0.19	95.20
26	-0.06	0.17	96.70	-0.05	0.18	96.90	0.03	0.19	95.70	0.02	0.18	95.90	0.02	0.18	95.70 **95.40**	0.02	0.18	95.80	-	-	-	0.03	0.19	95.60
27	-0.14	0.30	96.20	-0.14	0.30	96.20	0.04	0.42	95.40	0.03	0.42	95.50	0.01	0.37	95.30 **94.80**	0.01	0.37	95.30	-	-	-	0.04	0.42	95.40
28	-0.13	0.29	96.70	-0.13	0.29	96.80	0.05	0.40	96.40	0.04	0.40	96.40	0.02	0.36	96.30 **95.60**	0.02	0.36	96.30	-	-	-	0.05	0.40	96.30
29	-0.04	0.10	95.60	-0.04	0.10	96.10	0.01	0.11	95.20	-0.02	0.10	95.40	0.00	0.10	95.20 **95.20**	0.00	0.10	95.30	-	-	-	0.01	0.11	95.20
30	-0.03	0.10	96.00	-0.03	0.10	96.70	0.02	0.11	95.60	-0.01	0.10	95.50	0.01	0.10	95.60 **95.60**	0.01	0.10	95.90	-	-	-	0.02	0.11	95.60
31	-0.02	0.06	95.60	-0.02	0.06	96.00	0.00	0.06	94.70	-0.06	0.06	93.10	0.00	0.06	94.90 **94.80**	0.00	0.06	95.30	-	-	-	0.00	0.06	94.80
32	-0.01	0.06	95.50	-0.01	0.06	96.10	0.01	0.06	94.70	-0.05	0.06	93.90	0.01	0.06	94.70 **94.60**	0.01	0.06	95.30	-	-	-	0.01	0.06	94.60

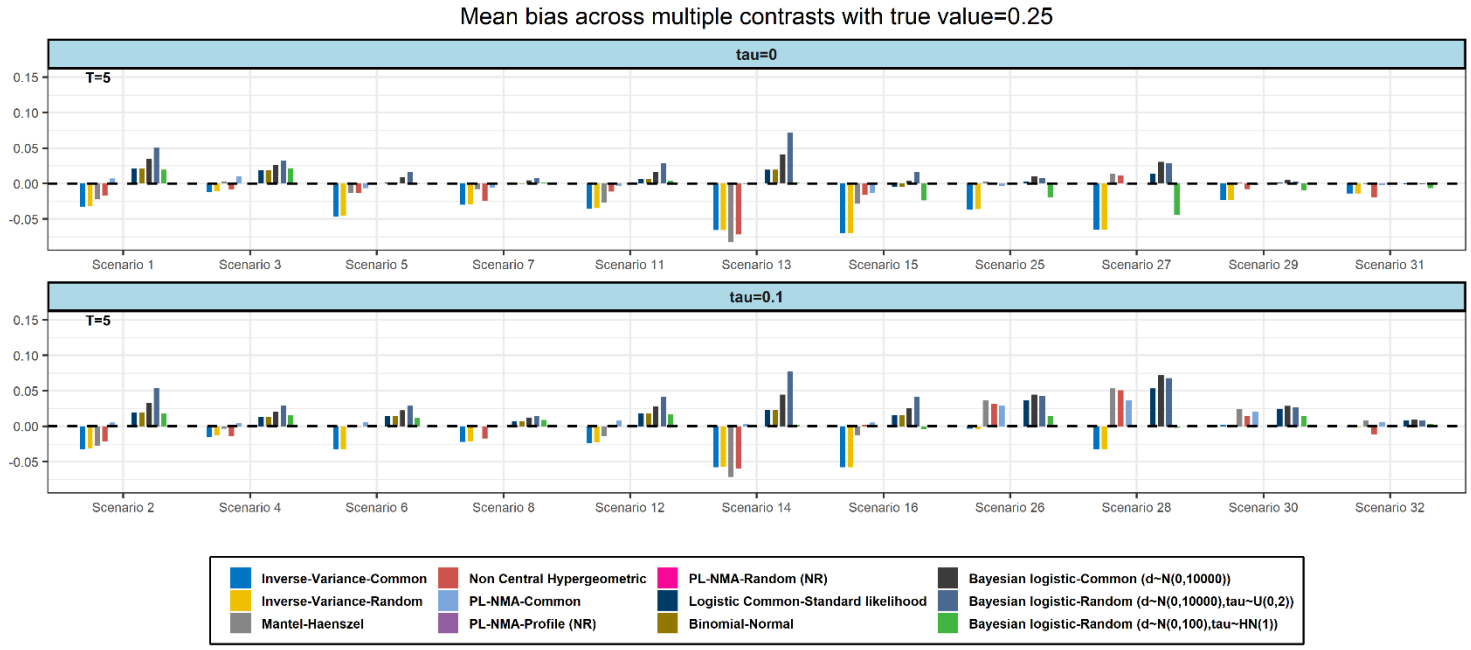
S-Table 1: Results in terms of mean bias, mean square error (MSE) and coverage probability (%) across all 32 scenarios and across Bayesian models.

#	Bayesian ($d \sim N(0, 100^2)$)			Bayesian ($d \sim N(0, 100^2), \tau \sim U(0, 2)$)			Bayesian ($d \sim N(0, 10^2), \tau \sim HN(1)$)		
	Mean bias	**MSE**	**Coverage (%)**	**Mean bias**	**MSE**	**Coverage (%)**	**Mean bias**	**MSE**	**Coverage (%)**
1	0.07	0.42	93.60	0.11	0.58	96.50	0.04	0.43	95.20
2	0.06	0.43	93.10	0.10	0.61	96.10	0.04	0.44	94.70
3	0.04	0.22	93.70	0.06	0.28	96.70	0.04	0.24	95.40
4	0.04	0.23	93.30	0.05	0.29	96.50	0.03	0.24	95.10
5	0.03	0.19	94.10	0.05	0.23	96.00	0.02	0.20	95.30
6	0.05	0.20	93.70	0.06	0.24	95.90	0.03	0.20	95.00
7	0.02	0.10	94.50	0.03	0.12	96.40	0.02	0.11	95.70
8	0.03	0.11	93.50	0.04	0.12	95.70	0.02	0.12	95.20
9	0.03	0.24	94.60	0.04	0.28	96.10	0.01	0.25	95.50
10	0.05	0.24	94.40	0.06	0.28	96.00	0.03	0.25	95.50
11	0.04	0.30	93.60	0.07	0.41	96.80	0.02	0.32	95.50
12	0.05	0.31	93.60	0.08	0.43	95.50	0.03	0.33	95.40
13	0.09	0.70	93.20	0.17	1.08	95.60	0.02	0.67	94.60
14	0.09	0.72	92.90	0.16	1.11	96.00	0.02	0.68	94.60
15	0.02	0.29	94.40	0.05	0.38	96.30	-0.03	0.29	95.20
16	0.05	0.30	93.70	0.09	0.40	95.80	0.00	0.30	95.00
17	0.05	0.20	94.70	0.07	0.26	97.30	0.03	0.20	96.40
18	0.02	0.18	95.20	0.05	0.24	97.30	0.01	0.19	96.40
19	0.10	0.45	94.30	0.15	0.61	97.00	0.03	0.41	96.00
20	0.06	0.41	94.90	0.11	0.57	97.00	0.01	0.39	95.90
21	0.03	0.11	95.70	0.04	0.13	97.80	0.02	0.11	97.20
22	0.01	0.10	95.00	0.02	0.13	97.70	0.01	0.11	97.20
23	0.02	0.06	94.70	0.03	0.07	97.60	0.02	0.07	96.70
24	0.01	0.06	95.20	0.01	0.07	97.70	0.01	0.06	97.00
25	0.03	0.21	94.50	0.03	0.23	96.20	0.00	0.20	95.60
26	0.05	0.20	95.00	0.05	0.23	96.50	0.02	0.20	95.80
27	0.07	0.48	93.50	0.08	0.55	95.60	0.00	0.43	94.90
28	0.09	0.45	94.50	0.10	0.52	96.30	0.01	0.41	95.80
29	0.02	0.11	94.80	0.02	0.12	96.50	0.00	0.11	96.10
30	0.03	0.11	94.80	0.03	0.12	96.90	0.02	0.12	96.70
31	0.01	0.06	94.50	0.01	0.07	96.50	0.00	0.06	96.30
32	0.01	0.06	94.20	0.01	0.07	96.40	0.01	0.07	96.40

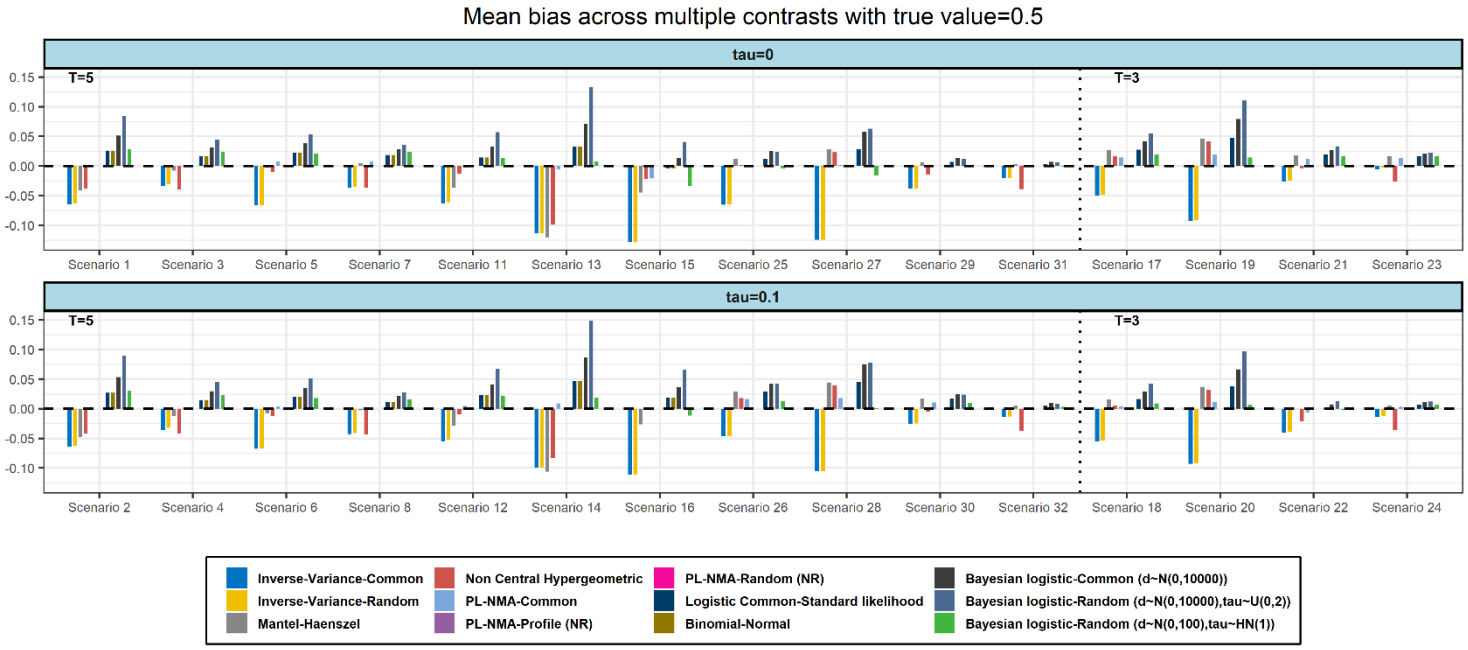
S-Table 2: Extent of studies with zero events in all treatment arms in the simulation study.

Scenario	Minimum	1st Quartile	Median	3rd Quartile	Maximum	Total number of studies per dataset
1	0	0	0	0	3	20
2	0	0	0	0	2	20
3	0	0	0	0	1	20
4	0	0	0	0	1	20
5	0	0	0	0	3	40
6	0	0	0	1	4	40
7	0	0	0	0	1	40
8	0	0	0	0	1	40
9	0	0	0	1	3	56
10	0	0	0	1	3	56
11	0	0	0	0	2	20
12	0	0	0	0	1	20
13	0	0	1	1	4	20
14	0	0	1	1	5	20
15	0	1	1	2	6	40
16	0	1	1	2	6	40
17	0	0	0	0	1	8
18	0	0	0	0	0	8
19	0	0	0	0	1	8
20	0	0	0	0	2	8
21	0	0	0	0	1	8
22	0	0	0	0	1	8
23	0	0	0	0	1	8
24	0	0	0	0	1	8
25	0	0	0	0	0	8
26	0	0	0	0	0	8
27	0	0	0	0	2	8
28	0	0	0	0	1	8
29	0	0	0	0	1	8
30	0	0	0	0	1	8
31	0	0	0	0	0	8
32	0	0	0	0	1	8
33	6	13	15	17	25	40

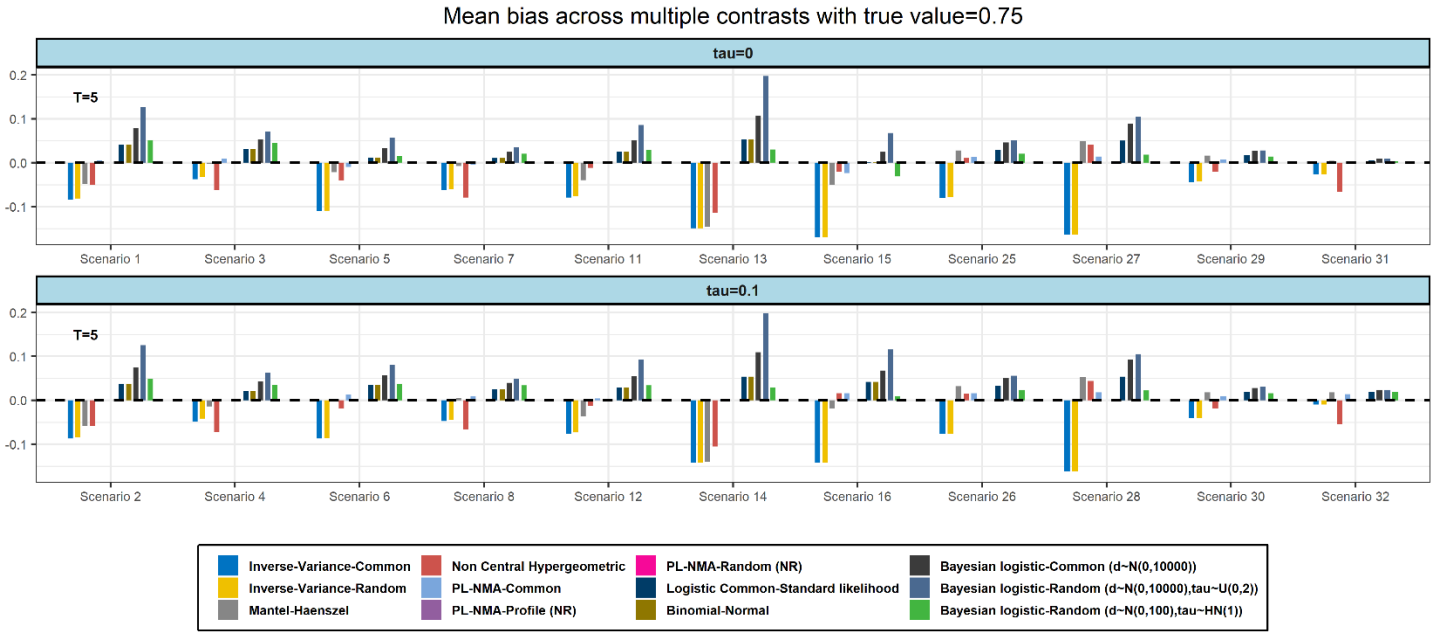
Results in terms of true $logORs$



S-Figure 2: Results in terms of bias when true $logOR$ is equal to 0.25. Models marked as *NR* are not relevant to the figure and thus no results are plotted.



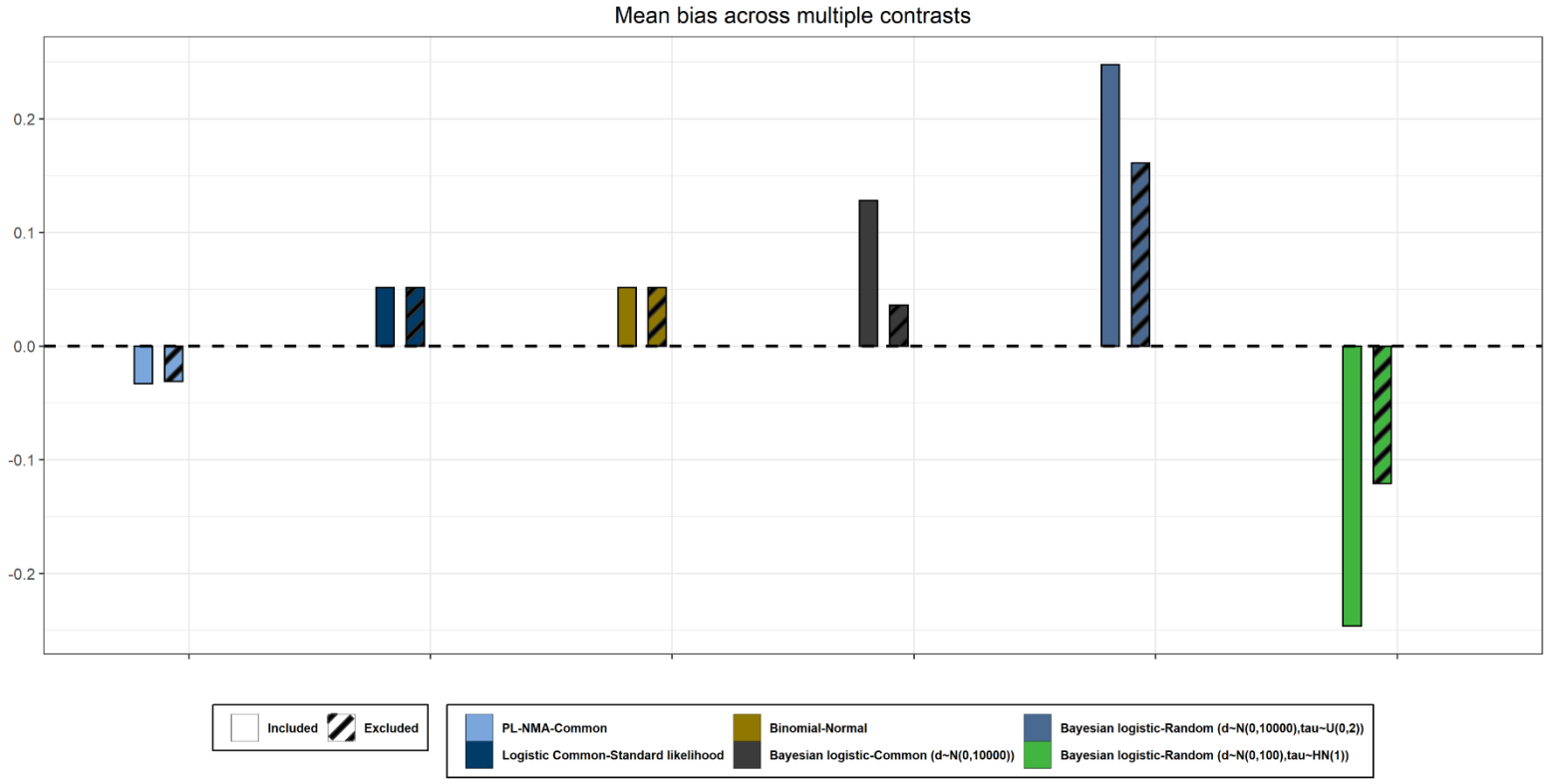
S-Figure 3: Results in terms of bias when true $logOR$ is equal to 0.50. Models marked as *NR* are not relevant to the figure and thus no results are plotted.



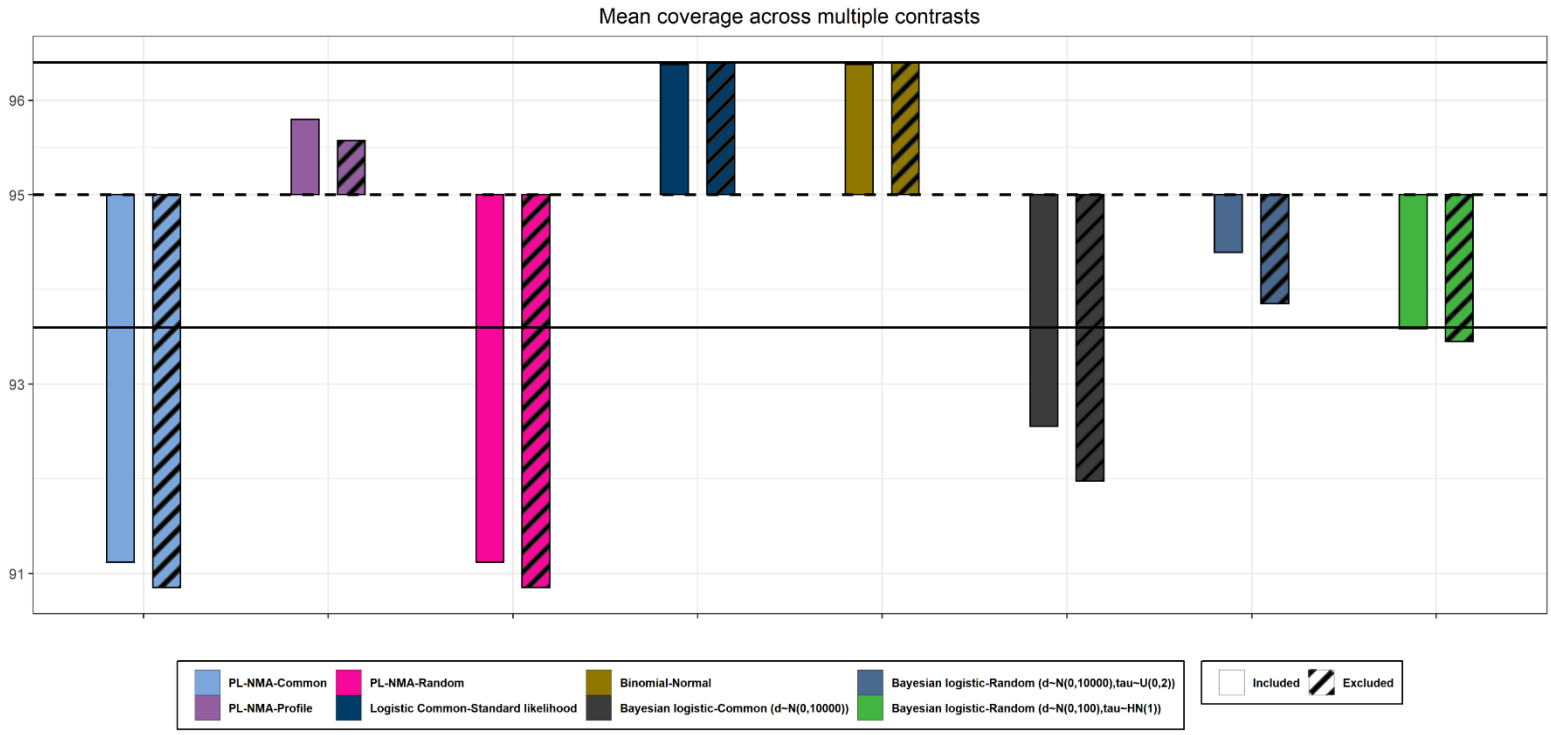
S-Figure 4: Results in terms of bias when true $logOR$ is equal to 0.75. Models marked as *NR* are not relevant to the figure and thus no results are plotted.

S-Figure 5*:* Results in terms of bias when true $logOR$ is equal to 1. Models marked as *NR* are not relevant to the figure and thus no results are plotted.

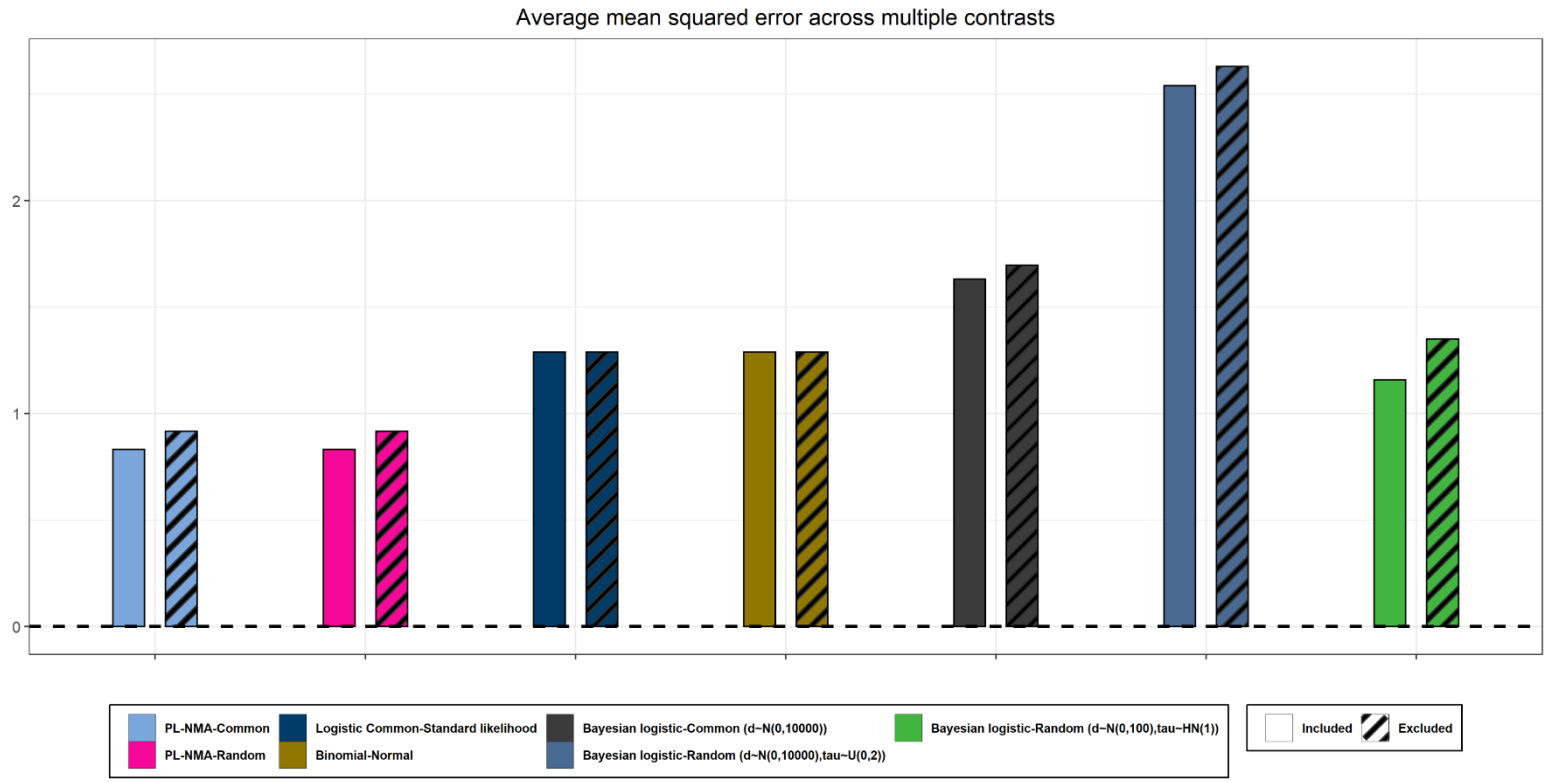
Results for Scenario 33



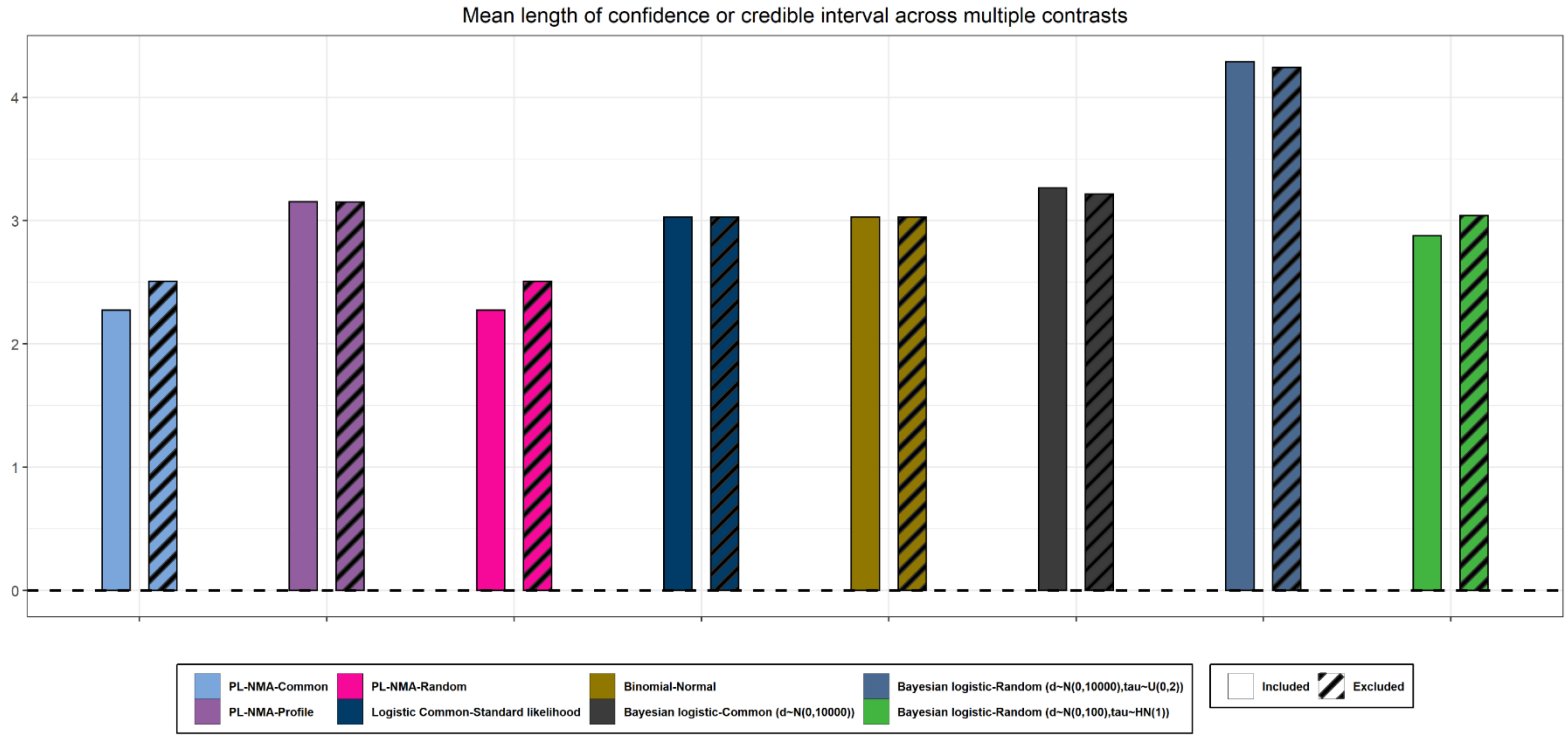
S-Figure 6: Mean bias across multiple contrasts for Scenario 33.



S-Figure 7: Mean coverage across multiple contrasts for Scenario 33.



S-Figure 8: Average mean squared error across multiple contrasts for Scenario 33.



S-Figure 9: Mean length of confidence or credible interval across multiple contrasts for Scenario 33.

Additional results for the clinical examples

Safety of inhaled medications for patients with chronic obstructive pulmonary disease

S-Table 4: Numerical results for the forest plot of the inhaled medications clinical example.

Model	ICS	LABA	LABA-ICS	TIO-HH	TIO-SMI
Mantel-Haenszel	1.03 [0.88,1.21]	0.93 [0.79,1.08]	0.79 [0.67,0.94]	0.92 [0.81,1.05]	1.52 [1.05,2.19]
Non Central Hypergeometric	1.02 [0.88,1.19]	0.94 [0.81,1.09]	0.81 [0.69, 0.95]	0.93 [0.82, 1.05]	1.50 [1.05, 2.14]
Penalized likelihood (Wald type, Profile likelihood)	1.03 [0.88, 1.21], [0.88 , 1.21]	0.93 [0.79, 1.09], [0.79, 1.09]	0.80 [0.67,0.94], [0.67, 0.94]	0.92 [0.81, 1.04], [0.81, 1.04]	1.50 [1.05, 2.16], [1.05, 2.15]
Logistic-Common Standard likelihood	1.03 [0.88, 1.21]	0.93 [0.79, 1.09]	0.79 [0.67, 0.94]	0.92 [0.81, 1.04]	1.51 [1.05, 2.17]
Bayesian logistic-Common (d~N(0,10000))	1.03 [0.88, 1.21]	0.93 [0.79, 1.09]	0.79 [0.67, 0.94]	0.92 [0.81, 1.04]	1.52 [1.06, 2.20]
Bayesian logistic-Random (d~N(0,10000),tau~U(0,2))	0.99 [0.77, 1.24]	0.95 [0.76, 1.21]	0.79 [0.62, 1.03]	0.92 [0.75, 1.15]	1.52 [1.00, 2.38]
Bayesian logistic-Random (d~N(0,100),tau~HN(1))	0.99 [0.76, 1.24]	0.95 [0.76, 1.21]	0.79 [0.62, 1.02]	0.92 [0.75, 1.15]	1.54 [1.01, 2.38]

Heterogeneity parameters

- **PL-NMA:** $\hat{\varphi} = 1$
- **Bayesian logistic-Random (d~N(0,10000), tau~U(0,2)):** $\hat{\tau}$ =0.12 [0.01, 0.34]
- **Bayesian logistic-Random (d~N(0,100), tau~HN(1)):** $\hat{\tau}$ =0.12 [0, 0.35]

Safety of different drug classes for chronic plaque psoriasis

A two step procedure for the exclusion of zero event studies applies for MH-NMA and NCH-NMA. At the first step all-zero event studies are excluded. After this step, the treatment design Anti-IL 23 vs. Placebo remains with studies that report zero events only for Placebo group. At a

second step MH-NMA and NCH-NMA are searching within each treatment design for the treatment arms for which all studies in the design report zero events and remove those treatment arms from the design. As a result, the second step leads to the further exclusion of the Placebo arms in the design Anti-IL 23 vs Placebo; this treatment design then remains with only two single arm studies and thus it is completely excluded from the analysis. The same two-stage procedure of exclusion applies to all treatment designs and leads to treatment designs Anti-IL 23 vs. Anti-IL 12/23 and Anti-IL 23 vs. Anti-TNF to be excluded from the analysis. All the treatment designs that involve the treatment Anti-IL 23 are needed to be removed and as a result the Anti-IL 23 nodes is discarded from the network.

S-Table 5: Numerical results for the forest plot of the psoriasis clinical example.

Model	Anti-IL 12/23	Anti-IL 17	Anti-IL 23	Anti-TNF	Apremilast
Mantel-Haenszel	1.23 [0.12, 12.50]	0.97 [0.14, 6.72]	No results	0.91 [0.14, 5.84]	0.37 [0.02, 7.02]
Non Central Hypereometric	1.28 [0.10, 16.17]	1.11 [0.19, 6.43]	No results	0.91 [0.14, 5.71]	0.41 [0.02, 6.63]
Penalized likelihood (Wald type, Profile likelihood)	1.30 [0.26, 6.62], [0.11, 19.55]	0.85 [0.29, 2.54], [0.22, 4.23]	2.54 [0.40, 16.09], [0.12, 72.87)	1.45 [0.46, 4.60], [0.19, 13.23]	0.43 [0.06, 2.91], [0.05, 4.83]
Logistic-Common Standard likelihood	1.34 [0.11, 15.98]	0.96 [0.21, 4.29]	5.72×10^{9} [0, Inf) (No results)	1.60 [0.26, 9.68]	0.41 [0.03, 6.64]
Bayesian logistic-Common (d~N(0,10000))	1.37 [0.10, 21.23]	1.08 [0.23, 6.18]	1.87×10^{4} [0.73, 1197×10^{15}] (No results)	1.80 [0.28, 14.90]	0.40 [0.01, 15.04]
Bayesian logistic-Random (d~N(0,10000),tau~U(0,2))	1.61 [0.08, 33.18]	1.35 [0.19, 12.63]	5.10×10^{5} [0.94, 7.83×10^{15}] (No results)	2.32 [0.25, 28.98]	0.38 [0.01, 23.49]
Bayesian logistic-Random (d~N(0,100),tau~HN(1))	0.66 [0.06, 7.81]	0.63 [0.12-2.85]	1.03 [0.09, 16.80]	0.74 [0.13, 4.23]	0.28 [0.01, 7.09]

Heterogeneity parameters

- **PL-NMA:** $\hat{\varphi} = 1$
- **Bayesian logistic-Random (d~N(0,10000), tau~U(0,2)):** $\hat{\tau}$ =0.97 [0.05, 1.94]
- **Bayesian logistic-Random (d~N(0,100), tau~HN(1)):** $\hat{\tau}$ =0.56 [0.03, 1.61]

Annex 2: Supplementary files for article in Part 3

Appendix 1: Questionnaire

Questionnaire for methodological project on the analysis of psychiatric symptom reduction by antipsychotics in children and adolescents

Short introduction to the project (please read this before starting answering the questions)

In this project we aim to estimate the effects of different antipsychotics for children and adolescents (Krause et al. 2018[1]). However, the available evidence for this population is very limited. To increase the confidence in our results, we plan to borrow information from available evidence on the same outcome for "general" patients (usually chronic adults with an exacerbation of positive symptoms, for example as those meta-analyzed by Huhn et al. 2019[2]) for whom more evidence is available. Thus, we need to estimate the 'difference' between the two populations. To do this properly, we would need your help by answering a questionnaire that we have prepared below.

More details about our methodology can be found in the following section "**Methodological details related to the project**".

Otherwise, please go directly to the questionnaire in the section "

Questions to be answered by experts" in page 4.

Methodological details related to the project

Sparse data are a common issue in comparative effectiveness research even at the meta-analysis level. We use as an exemplar sparse network, the network formed by 19 RCTs comparing 14 antipsychotics and placebo for the reduction of overall psychiatric symptoms (PANSS total score) in children and adolescents. This network is depicted in **Figure 1**. The 19 RCTs provide 21 out of the 105 (20%) theoretically possible direct comparisons (lines in the graph). Performing a 'standard' network meta-analysis (NMA) in this network, which combines direct (example: all studies comparing aripiprazole and paliperidone with each other) and indirect evidence (example: estimating the difference between aripiprazole and paliperidone indirectly from aripiprazole versus placebo and paliperidone versus placebo), is possibly of limited validity. Specifically, the resulting relative effects are equally or even less precise for some comparisons than those obtained from pairwise meta-analysis and the model assumptions cannot be evaluated. However, the pace of new research production for this network is very slow and it is of great importance to use any available piece of information to identify the most beneficial drugs for children and adolescents.

The main idea behind the following questionnaire is that we could use external evidence from a dense (i.e. informative) network after some form of 'adaptation' of the results to the population of interest and increase the amount of information for the sparse network for children and adolescents. In other words, we could use this adapted external evidence as 'prior' information in a Bayesian NMA model for children and adolescents. In this way, we aim to obtain more precise and more robust (reliable) results for the comparative efficacy of the drugs for children and adolescents.

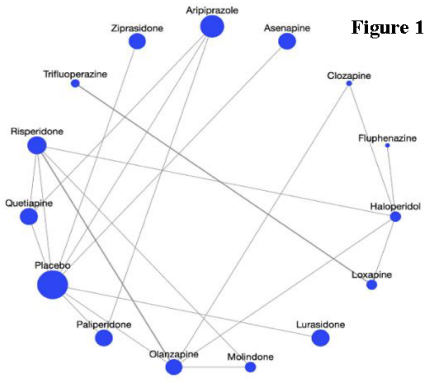
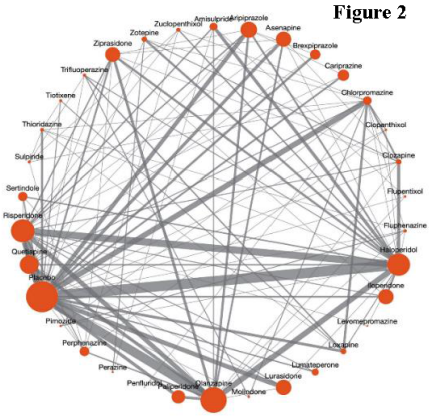


Figure 1

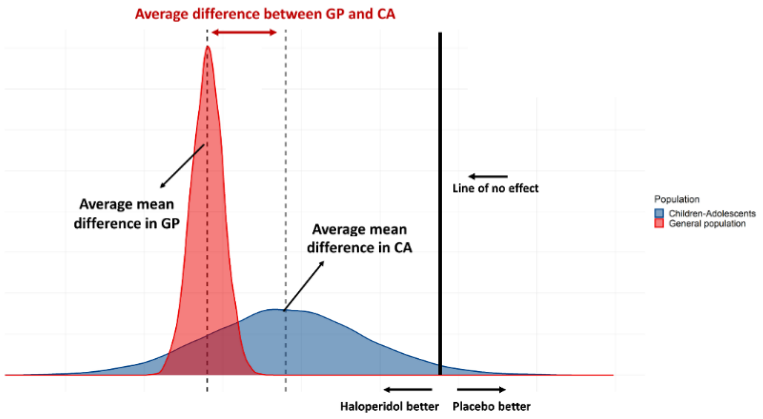
The external informative network we consider here is a network evaluating the drugs for the same outcome (overall symptoms) in "general" adult patients with schizophrenia (**Figure 2**) as they were, for example, meta-analyzed in Huhn et al. 2019[2]. These studies, albeit they do not always use exactly the same definitions, can be characterised as usually chronic patients with an acute exacerbation of positive symptoms. This network consists of 255 studies and 116 direct comparisons. Note that we will only use the evidence from this network for the drugs already existing in the children and adolescent network; hence we will not draw any inference for drugs that have never been used for the population of interest. Of course, some differences among the two populations are inevitable and, therefore, we cannot use the results from the general population without incorporating these differeces. Since we only have aggregate (i.e. summary) data from the studies, we will rely on the average difference between the outcome in the two populations for each drug.

Figure 2



From a statistical standpoint, the average difference in the outcome between the two populations is represented by the difference of the mean of the respective distributions. **Figure 3** shows a theoretical example about the difference of the means of the two distributions for the comparison haloperidol vs placebo. Such differences for all comparisons can be obtained either using the data (which we will do, as well) or using expert opinion (this is what we need your help for). Then, these differences will be modelled appropriately to obtain NMA results from the general patient data applicable to children and adolescents which, in turn, will be used as prior information in the NMA for children and adolescents.

Figure 3



References

1. Krause M, Zhu Y, Huhn M, Schneider-Thoma J, Bighelli I, Chaimani A, Leucht S. Efficacy, acceptability, and tolerability of antipsychotics in children and adolescents with schizophrenia: A network meta-analysis. Eur Neuropsychopharmacol. 2018 Jun;28(6):659-674

2. Huhn M, Nikolakopoulou A, Schneider-Thoma J, Krause M, Samara M, Peter N, Arndt T, Bäckers L, Rothe P, Cipriani A, Davis J, Salanti G, Leucht S. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. Lancet. 2019 Sep 14;394(10202):939-951.

Questions to be answered by experts

Q1: Column 2 of the next table shows a theoretical reduction (change score) in PANSS total score for overall symptoms for the antipsychotics of Figure 2 in "general patients" (chronic adults with acute exacerbation of positive symptoms).

Please fill into column 3 how much you expect that the corresponding reductions would be for children and adolescents. Please, also provide a standard deviation for the value you provide in column 4.

You may skip the antipsychotics which you are not familiar with.

Please also tick the box in column 5 regarding how you think the reduction of the PANSS total score in children and adolescents qualitatively differs from that in general patients.

The first row in red presents a fictional example: If the PANSS total score reduction from baseline to endpoint for "general patients" for a given fictional antipsychotic X is -26 then a possible answer for children and adolescents may be:

- Expected PANSS total score reduction from baseline to endpoint in Children-Adolescents=-29 meaning three points more improvement
- Standard Deviation for the expected PANSS reduction in Children-Adolescents =±10
- Expected treatment response in qualitative terms in Children-Adolescents compared to "general patients" = A bit better

1. Drug	2. PANSS total score reduction from baseline to endpoint for "general adult patients" (chronic adults with acute exacerbation of positive symptoms)	3. Expected PANSS total score reduction from baseline to endpoint in Children-Adolescents	4. Standard deviation for the expected PANSS reduction in Children-Adolescents	5. Expected treatment response in qualitative terms in Children-Adolescents compared to "general adult patients"
Fictional antipsychotic X	-26	-29	±10	☐ Much better ☒ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Clozapine	-27.8			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Olanzapine	-21.2			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse

Risperidone	-21			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Paliperidone	-19.8			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Haloperidol	-19.4			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Loxapine	-19			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Quetiapine	-18.4			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Molindone	-18.4			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Aripiprazole	-18.2			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Ziprasidone	-18.2			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Asenapine	-17.8			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Lurasidone	-17.2			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Fluphenazine	-14.8			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse

Trifluoperazine	-14.8			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse
Placebo	-10			☐ Much better ☐ A bit better ☐ Same ☐ A bit worse ☐ Much worse

Q2: In a scale from 1 to 10 how confident you are in the numbers that you gave in **Q1**? (1=not confident, 10=very confident)

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Q3: What are the most probable population related factors that may differentiate the effectiveness of the antipsychotics between general patients and children and adolescents?

Q4: Consider that a new drug with very positive results for the general population has come out (very efficacious without important side effects) but not tested in children and adolescents. Would you consider it for treating children/adolescents with schizophrenia? If yes, in which situations (please fill-in the box below)? **Please disregard any legal concerns with off-label use.**

Certainly yes ☐

Possibly yes ☐

Maybe ☐

Possibly no ☐

Certainly no ☐

If not, why would you <u>not</u> consider giving the potential new drug to children and adolescents when evidence is available only for the general population?

Q5: Your main work is

Research ☐

Clinical practice ☐

Both ☐

Q6: How many years have you been working in the field of schizophrenia?

Q7: You have experience with

Child psychiatry ☐

Adult psychiatry ☐

Both ☐

Q8: What kind of experience do you have with studies about antipsychotics?

Involved in randomized clinical trials ☐

Involved in meta-analyses of clinical trials ☐

Involved in observational studies ☐

Any other study related experience ☐

Appendix 2: Consistency checks, Ranking and League tables

Appendix Table 3: Consistency checks for the naive synthesis model.

Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.39 [-0.50, -0.27]	-0.46 [-0.56, -0.36]	0.07 [-0.07, 0.22]	0.84	0.31
Olanzapine vs Placebo	-0.50 [-0.58, -0.42]	-0.59 [-0.66, -0.51]	0.09 [-0.02, 0.20]	0.94	0.12
Paliperidone vs Placebo	-0.52 [-0.65, -0.39]	-0.53 [-0.66, -0.40]	0.01 [-0.19, 0.21]	0.51	0.98
Quetiapine vs Placebo	-0.30 [-0.41, -0.19]	-0.41 [-0.52, -0.30]	0.11 [-0.05, 0.26]	0.90	0.20
Risperidone vs Placebo	-0.44 [-0.52, -0.35]	-0.50 [-0.59, -0.41]	0.06 [-0.06, 0.19]	0.85	0.29

Appendix Table 4: Consistency checks for the NMA model with informative priors obtained from GP using a data based approach for β and no downweight.

Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.40 [-0.60, -0.19]	-0.50 [-0.82, -0.18]	0.10 [-0.27, 0.47]	0.71	0.57
Olanzapine vs Placebo	-0.58 [-0.80, -0.37]	-0.73 [-1.05, -0.39]	0.15 [-0.27, 0.55]	0.77	0.46
Paliperidone vs Placebo	-0.45 [-0.71, -0.22]	-0.34 [-0.76, 0.09]	-0.11 [-0.62, 0.36]	0.32	0.64
Quetiapine vs Placebo	-0.41 [-0.63, -0.17]	-0.33 [-0.74, 0.11]	-0.07 [-0.55, 0.41]	0.38	0.76
Risperidone vs Placebo	-0.57 [-0.80, -0.32]	-0.49 [-0.85, -0.14]	-0.08 [-0.50, 0.36]	0.35	0.71

Appendix Table 5: Consistency checks for the NMA model with informative priors obtained from GP using a data based approach for β and moderate downweight to all GP studies with high RoB.

Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.39 [-0.60, -0.20]	-0.50 [-0.80, -0.19]	0.10 [-0.27, 0.46]	0.71	0.57
Olanzapine vs Placebo	-0.58 [-0.79, -0.38]	-0.73 [-1.09, -0.38]	0.15 [-0.26, 0.55]	0.77	0.46
Paliperidone vs Placebo	-0.45 [-0.67, -0.23]	-0.35 [-0.74, 0.05]	-0.10 [-0.56, 0.35]	0.32	0.64
Quetiapine vs Placebo	-0.42 [-0.64, -0.18]	-0.33 [-0.71, 0.06]	-0.09 [-0.54, 0.36]	0.35	0.71
Risperidone vs Placebo	-0.57 [-0.80, -0.31]	-0.50 [-0.86, -0.14]	-0.06 [-0.51, 0.39]	0.38	0.76

Appendix Table 6: Consistency checks for the NMA model with informative priors obtained from GP using a data based approach for β and moderate downweight to all GP studies with interventions in $T_a - T_c$.

Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.41 [-0.57, -0.23]	-0.51 [-0.81, -0.20]	0.10 [-0.26, 0.44]	0.73	0.54
Olanzapine vs Placebo	-0.59 [-0.76, -0.41]	-0.73 [-1.07, -0.39]	0.14 [-0.23, 0.53]	0.76	0.48
Paliperidone vs Placebo	-0.46 [-0.66, -0.26]	-0.36 [-0.74, 0.04]	-0.10 [-0.54, 0.32]	0.31	0.62
Quetiapine vs Placebo	-0.43 [-0.64, -0.22]	-0.35 [-0.75, 0.05]	-0.08 [-0.53, 0.36]	0.36	0.71
Risperidone vs Placebo	-0.54 [-0.74, -0.32]	-0.49 [-0.84, -0.12]	-0.05 [-0.46, 0.36]	0.41	0.81

Appendix Table 7: Consistency checks for the NMA model with informative priors obtained from GP using expert's opinion for β and no downweight.

Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.41 [-0.65, -0.19]	-0.47 [-0.80, -0.14]	0.06 [-0.35, 0.47]	0.62	0.77
Olanzapine vs Placebo	-0.58 [-0.83, -0.33]	-0.69 [-1.04, -0.33]	0.11 [-0.31, 0.54]	0.71	0.58
Paliperidone vs Placebo	-0.47 [-0.74, -0.21]	-0.34 [-0.75, 0.08]	-0.13 [-0.62, 0.35]	0.30	0.59
Quetiapine vs Placebo	-0.33 [-0.57, -0.07]	-0.37 [-0.75, 0.06]	0.04 [-0.45, 0.52]	0.55	0.90
Risperidone vs Placebo	-0.66 [-0.88, -0.42]	-0.42 [-0.77, -0.06]	-0.24 [-0.68, 0.19]	0.15	0.29

Appendix Table 8: Consistency checks for the NMA model with informative priors obtained from GP using expert's opinion for β and moderate downweight to all GP studies with high RoB.

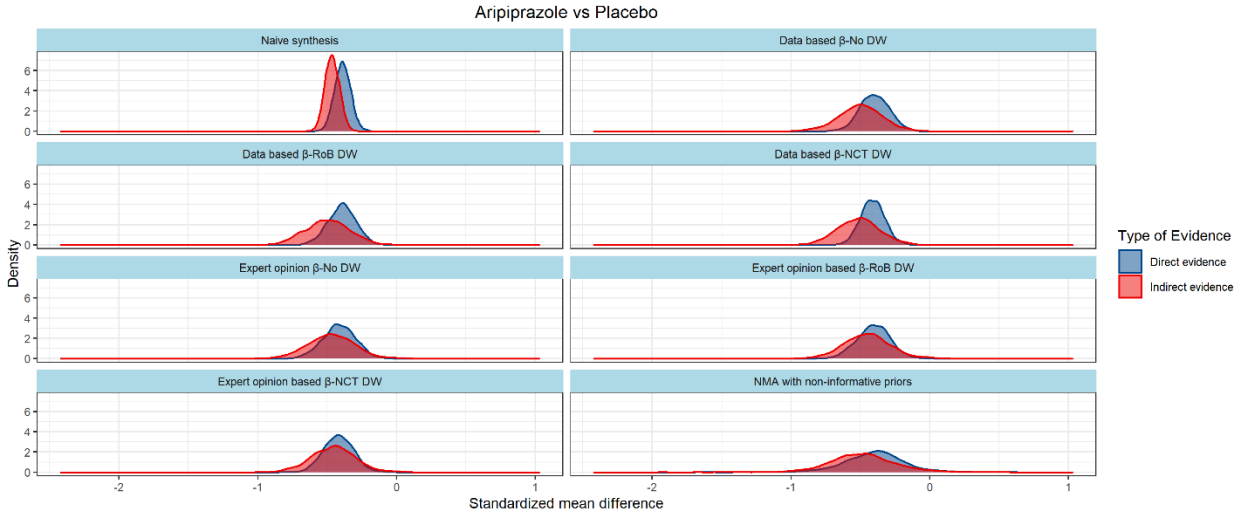
Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.41 [-0.64, -0.17]	-0.45 [-0.77, -0.10]	0.05 [-0.37, 0.44]	0.59	0.81
Olanzapine vs Placebo	-0.57 [-0.81, -0.32]	-0.76 [-1.11, -0.41]	0.20 [-0.22, 0.63]	0.81	0.37
Paliperidone vs Placebo	-0.47 [-0.72, -0.20]	-0.32 [-0.73, 0.07]	-0.14 [-0.62, 0.33]	0.27	0.55
Quetiapine vs Placebo	-0.28 [-0.52, -0.04]	-0.35 [-0.79, 0.07]	0.08 [-0.43, 0.58]	0.61	0.78
Risperidone vs Placebo	-0.63 [-0.85, -0.39]	-0.45 [-0.80, -0.07]	-0.18 [-0.60, 0.24]	0.20	0.40

Appendix Table 9: Consistency checks for the NMA model with informative priors obtained from GP using expert's opinion for β and moderate downweight to all GP studies with interventions in $T_a - T_c$.

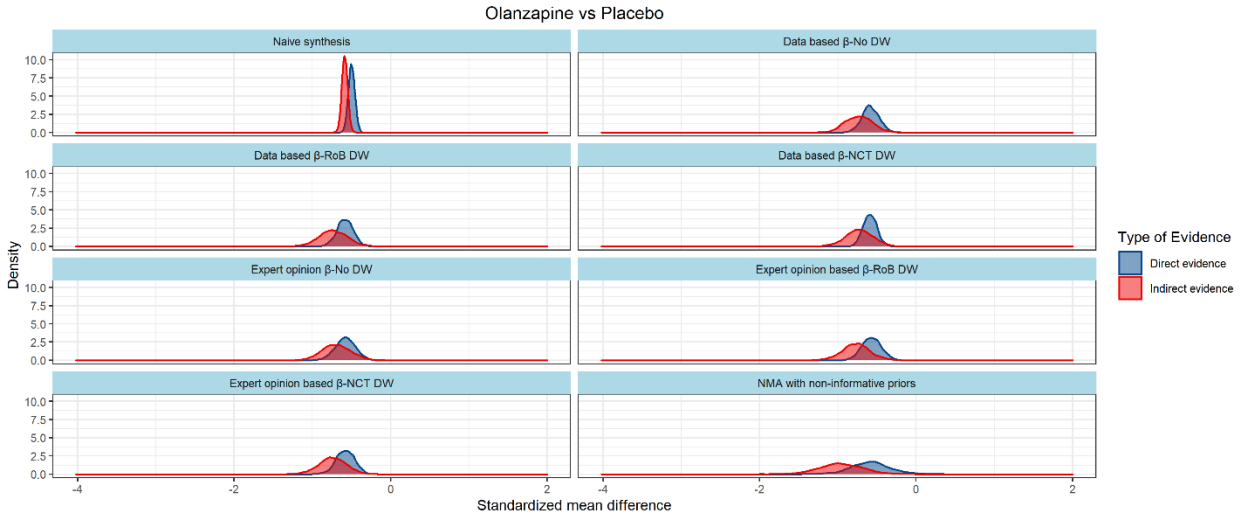
Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.42 [-0.62, -0.21]	-0.46 [-0.77, -0.13]	0.04 [-0.34, 0.41]	0.59	0.82
Olanzapine vs Placebo	-0.58 [-0.82, -0.36]	-0.74 [-1.09, -0.40]	0.16 [-0.25, 0.58]	0.78	0.44
Paliperidone vs Placebo	-0.48 [-0.73, -0.24]	-0.33 [-0.71, 0.08]	-0.16 [-0.65, 0.29]	0.25	0.49
Quetiapine vs Placebo	-0.27 [-0.48, -0.04]	-0.36 [-0.78, 0.07]	0.10 [-0.38, 0.56]	0.65	0.69
Risperidone vs Placebo	-0.62 [-0.83, -0.39]	-0.44 [-0.81, -0.08]	-0.18 [-0.60, 0.25]	0.20	0.40

Appendix Table 10: Consistency checks for the NMA model with non-informative priors.

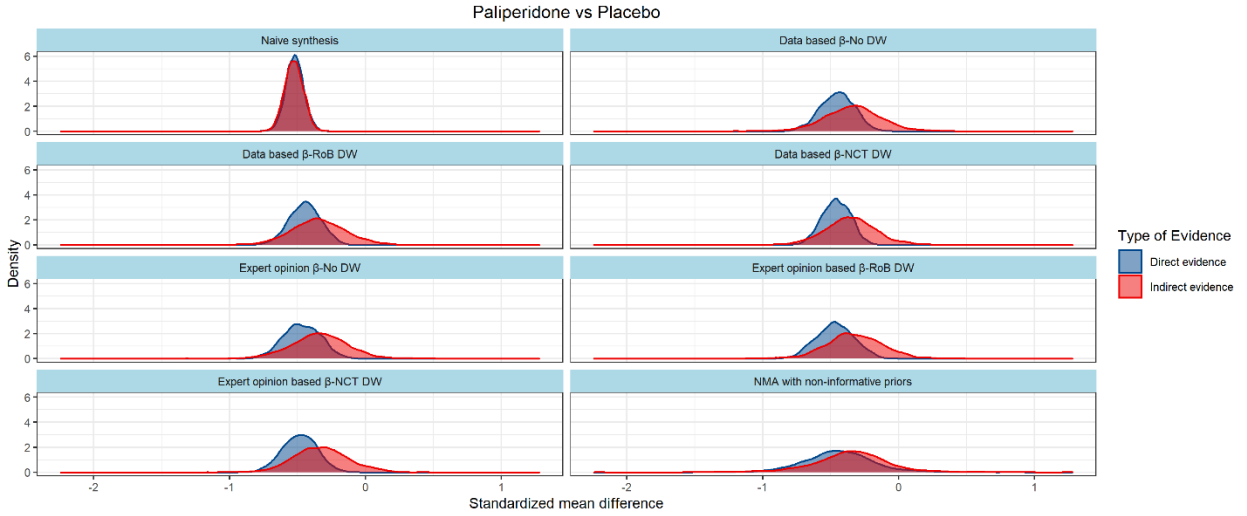
Comparison	Direct [95% CI]	Indirect [95% CI]	Difference [95% CI]	P(diff>0)	p-value
Aripiprazole vs Placebo	-0.40 [-0.86, 0.07]	-0.49 [-0.95, 0.02]	0.09 [-0.62, 0.76]	0.63	0.74
Olanzapine vs Placebo	-0.58 [-1.10, -0.05]	-0.96 [-1.51, -0.39]	0.38 [-0.43, 1.19]	0.85	0.30
Paliperidone vs Placebo	-0.44 [-0.95, 0.12]	-0.35 [-1.02, 0.27]	-0.09 [-0.85, 0.71]	0.40	0.79
Quetiapine vs Placebo	-0.40 [-0.88, 0.07]	-0.37 [-0.92, 0.19]	-0.02 [-0.75, 0.69]	0.47	0.95
Risperidone vs Placebo	-0.80 [-1.21, -0.38]	-0.49 [-1.05, 0.03]	-0.32 [-0.99, 0.34]	0.15	0.31



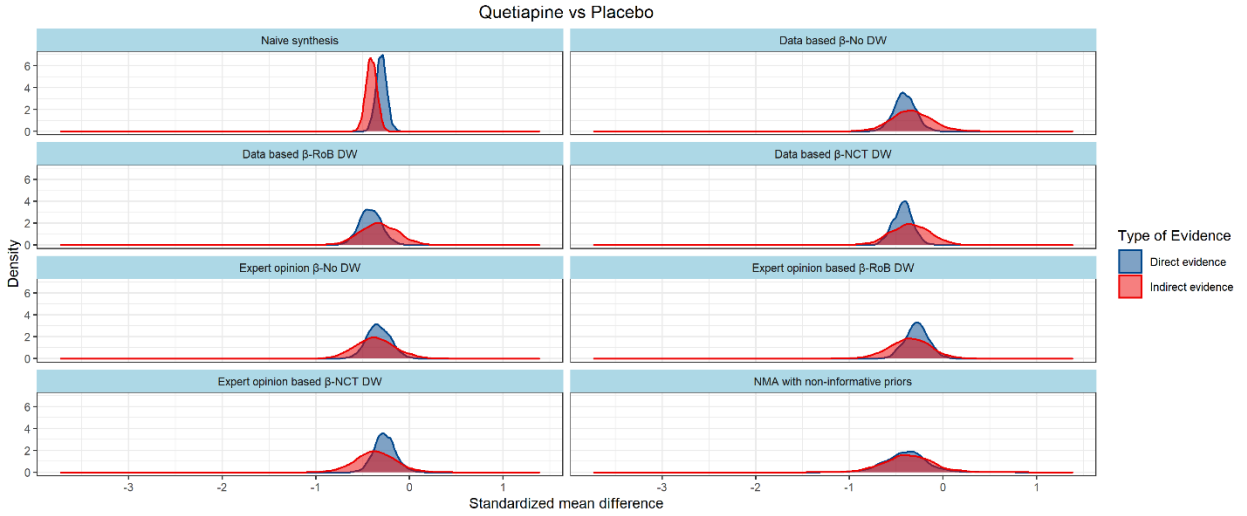
Appendix Figure 10: Posterior densities for direct and indirect evidence in terms of comparison Aripiprazole versus Placebo across all the different models.



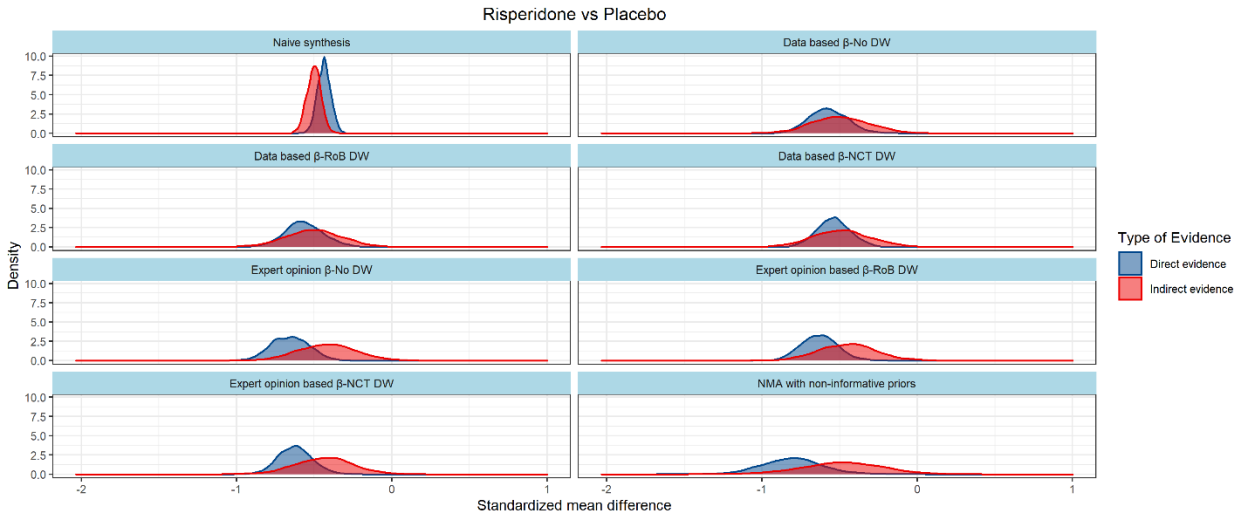
Appendix Figure 11: Posterior densities for direct and indirect evidence in terms of comparison Olanzapine versus Placebo across all the different models.



Appendix Figure 12: Posterior densities for direct and indirect evidence in terms of comparison Paliperidone versus Placebo across all the different models.

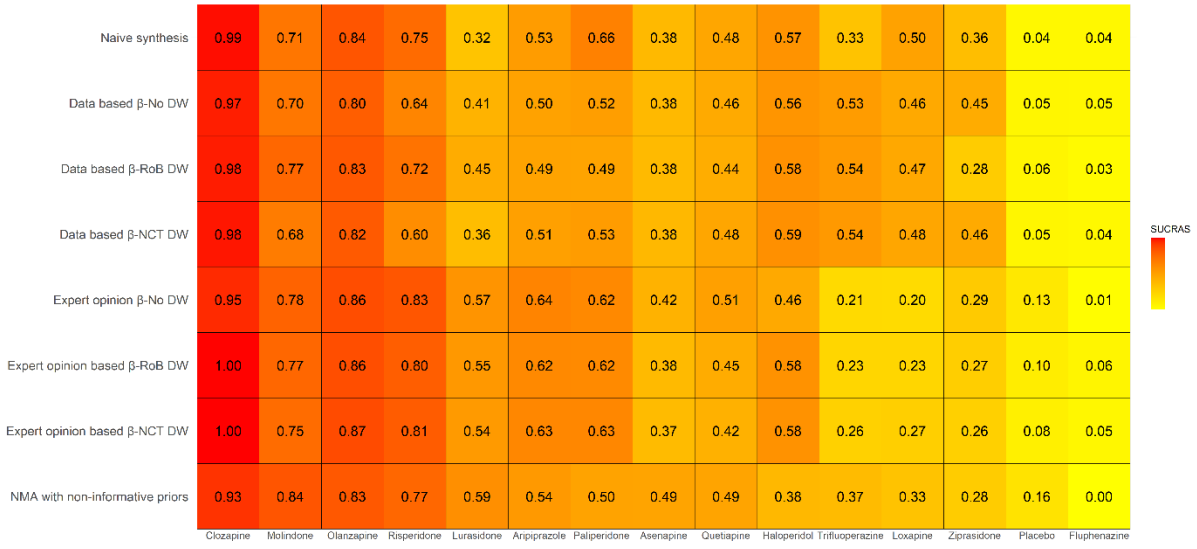


Appendix Figure 13: Posterior densities for direct and indirect evidence in terms of comparison Quetiapine versus Placebo across all the different models.



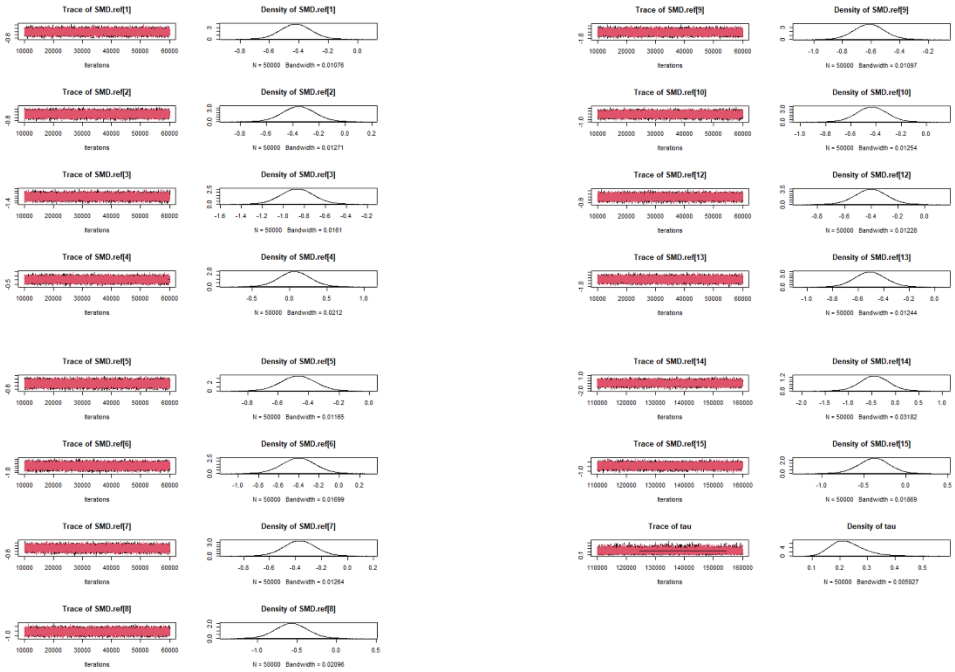
Appendix Figure 14: Posterior densities for direct and indirect evidence in terms of comparison Risperidone versus Placebo across all the different models.

Ranking between treatments across all models

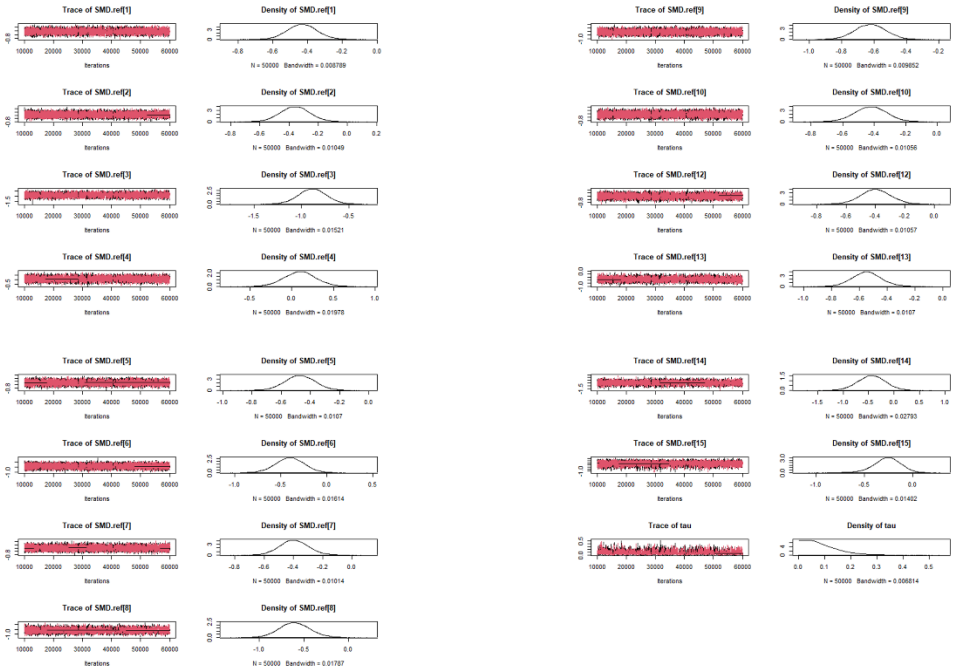


Appendix Figure 15: Ranking between all the drugs (14 antipsychotics and Placebo) as obtained from each one of the different models.

Traceplots



Appendix Figure 16: Trace plots for the NMA model with informative priors obtained from GP using a data based approach for β and no downweight.



Appendix Figure 17: Trace plots for the NMA model with informative priors obtained from GP using a data based approach for β and moderate downweight to all GP studies with high RoB

Appendix Figure 18: Trace plots for the NMA model with informative priors obtained from GP using a data based approach for β and moderate downweight to all GP studies with interventions in T_a-T_c

Appendix Figure 19: Trace plots for the NMA model with informative priors obtained from GP expert's opinion for β and no downweight.

Appendix Figure 20: Trace plots for the NMA model with informative priors obtained from GP expert's opinion for β and moderate downweight to all GP studies with high RoB

Appendix Figure 21: Trace plots for the NMA model with informative priors obtained from GP expert's opinion for β and moderate downweight to all GP studies with interventions in T_a-T_c

League tables

Appendix Table 9: League table for all comparisons in CA network using a naive synthesis NMA model

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.065(-0.182,0.057)	0.441(0.266,0.628)	-0.439(-0.753,-0.134)	0.02(-0.068,0.107)	-0.115(-0.329,0.093)	-0.088(-0.214,0.03)	0.128(-0.182,0.446)	0.148(0.061,0.234)	0.053(-0.064,0.17)	-0.431(-0.513,-0.352)	-0.019(-0.121,0.089)	0.092(-0.001,0.183)	-0.158(-0.427,0.127)	-0.074(-0.184,0.036)
Asenapine	0.065(-0.057,0.182)	NA	0.505(0.318,0.691)	-0.375(-0.689,-0.061)	0.084(-0.019,0.186)	-0.05(-0.279,0.166)	-0.024(-0.152,0.099)	0.192(-0.115,0.526)	0.213(0.116,0.311)	0.117(-0.016,0.245)	-0.366(-0.458,-0.275)	0.046(-0.07,0.164)	0.156(0.05,0.264)	-0.094(-0.373,0.181)	-0.009(-0.132,0.109)
Clozapine	-0.441(-0.628,-0.266)	-0.505(-0.691,-0.318)	NA	-0.88(-1.214,-0.534)	-0.421(-0.597,-0.259)	-0.556(-0.815,-0.286)	-0.529(-0.714,-0.345)	-0.313(-0.657,0.025)	-0.292(-0.463,-0.117)	-0.388(-0.574,-0.199)	-0.872(-1.046,-0.703)	-0.459(-0.647,-0.276)	-0.349(-0.526,-0.172)	-0.599(-0.913,-0.277)	-0.515(-0.696,-0.337)
Fluphenazine	0.439(0.134,0.753)	0.375(0.061,0.689)	0.88(0.534,1.214)	NA	0.459(0.156,0.759)	0.324(-0.04,0.698)	0.351(0.038,0.666)	0.567(0.131,0.989)	0.588(0.288,0.893)	0.492(0.171,0.807)	0.008(-0.303,0.307)	0.421(0.111,0.735)	0.531(0.221,0.833)	0.281(-0.127,0.676)	0.365(0.055,0.674)
Haloperidol	-0.02(-0.107,0.068)	-0.084(-0.186,0.019)	0.421(0.259,0.597)	-0.459(-0.759,-0.156)	NA	-0.135(-0.342,0.072)	-0.108(-0.212,-0.005)	0.108(-0.192,0.422)	0.129(0.064,0.197)	0.033(-0.076,0.138)	-0.45(-0.51,-0.391)	-0.038(-0.123,0.049)	0.072(0.001,0.139)	-0.178(-0.43,0.09)	-0.093(-0.184,-0.004)
Loxapine	0.115(-0.093,0.329)	0.05(-0.166,0.279)	0.556(0.286,0.815)	-0.324(-0.698,0.04)	0.135(-0.072,0.342)	NA	0.027(-0.202,0.25)	0.243(-0.1,0.583)	0.263(0.056,0.476)	0.168(-0.045,0.397)	-0.316(-0.518,-0.11)	0.096(-0.114,0.304)	0.207(-0.003,0.414)	-0.043(-0.292,0.186)	0.041(-0.187,0.248)
Lurasidone	0.088(-0.03,0.214)	0.024(-0.099,0.152)	0.529(0.345,0.714)	-0.351(-0.666,-0.038)	0.108(0.005,0.212)	-0.027(-0.25,0.202)	NA	0.216(-0.091,0.541)	0.237(0.138,0.338)	0.141(0.018,0.266)	-0.343(-0.428,-0.256)	0.07(-0.045,0.182)	0.18(0.08,0.28)	-0.07(-0.333,0.211)	0.014(-0.109,0.136)
Molindone	-0.128(-0.446,0.182)	-0.192(-0.526,0.115)	0.313(-0.025,0.657)	-0.567(-0.989,-0.131)	-0.108(-0.422,0.192)	-0.243(-0.583,0.1)	-0.216(-0.541,0.091)	NA	0.021(-0.286,0.316)	-0.075(-0.407,0.231)	-0.559(-0.874,-0.261)	-0.146(-0.467,0.167)	-0.036(-0.34,0.263)	-0.286(-0.627,0.068)	-0.202(-0.515,0.117)
Olanzapine	-0.148(-0.234,-0.061)	-0.213(-0.311,-0.116)	0.292(0.117,0.463)	-0.588(-0.893,-0.288)	-0.129(-0.197,-0.064)	-0.263(-0.476,-0.056)	-0.237(-0.338,-0.138)	-0.021(-0.316,0.286)	NA	-0.096(-0.196,-0.002)	-0.579(-0.635,-0.525)	-0.167(-0.251,-0.08)	-0.057(-0.123,0.014)	-0.307(-0.563,-0.037)	-0.222(-0.313,-0.136)
Paliperidone	-0.053(-0.17,0.064)	-0.117(-0.245,0.016)	0.388(0.199,0.574)	-0.492(-0.807,-0.171)	-0.033(-0.138,0.076)	-0.168(-0.397,0.045)	-0.141(-0.266,-0.018)	0.075(-0.231,0.407)	0.096(0.002,0.196)	NA	-0.484(-0.578,-0.386)	-0.071(-0.19,0.05)	0.039(-0.066,0.147)	-0.211(-0.497,0.074)	-0.127(-0.246,0.003)
Placebo	0.431(0.352,0.513)	0.366(0.275,0.458)	0.872(0.703,1.046)	-0.008(-0.307,0.303)	0.45(0.391,0.51)	0.316(0.11,0.518)	0.343(0.256,0.428)	0.559(0.261,0.874)	0.579(0.525,0.635)	0.484(0.386,0.578)	NA	0.412(0.337,0.494)	0.523(0.463,0.585)	0.273(0.014,0.547)	0.357(0.267,0.445)
Quetiapine	0.019(-0.089,0.121)	-0.046(-0.164,0.07)	0.459(0.276,0.647)	-0.421(-0.735,-0.111)	0.038(-0.049,0.123)	-0.096(-0.304,0.114)	-0.07(-0.182,0.045)	0.146(-0.167,0.467)	0.167(0.08,0.251)	0.071(-0.05,0.19)	-0.412(-0.494,-0.337)	NA	0.11(0.023,0.197)	-0.14(-0.41,0.137)	-0.055(-0.165,0.054)
Risperidone	-0.092(-0.183,0.001)	-0.156(-0.264,-0.05)	0.349(0.172,0.526)	-0.531(-0.833,-0.221)	-0.072(-0.139,-0.001)	-0.207(-0.414,0.003)	-0.18(-0.28,-0.08)	0.036(-0.263,0.34)	0.057(-0.014,0.123)	-0.039(-0.147,0.066)	-0.523(-0.585,-0.463)	-0.11(-0.197,-0.023)	NA	-0.25(-0.51,0.026)	-0.166(-0.26,-0.074)
Trifluoperazine	0.158(-0.127,0.427)	0.094(-0.181,0.373)	0.599(0.277,0.913)	-0.281(-0.676,0.127)	0.178(-0.09,0.43)	0.043(-0.186,0.292)	0.07(-0.211,0.333)	0.286(-0.068,0.627)	0.307(0.037,0.563)	0.211(-0.074,0.497)	-0.273(-0.547,-0.014)	0.14(-0.137,0.41)	0.25(-0.026,0.51)	NA	0.084(-0.204,0.352)
Ziprasidone	0.074(-0.036,0.184)	0.009(-0.109,0.132)	0.515(0.337,0.696)	-0.365(-0.674,-0.055)	0.093(0.004,0.184)	-0.041(-0.248,0.187)	-0.014(-0.136,0.109)	0.202(-0.117,0.515)	0.222(0.136,0.313)	0.127(-0.003,0.246)	-0.357(-0.445,-0.267)	0.055(-0.054,0.165)	0.166(0.074,0.26)	-0.084(-0.352,0.204)	NA

Appendix Table 10: League table for all comparisons in CA network. Informative priors from GP studies using a data based approach for β and no downweight for GP studies.

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.07(-0.377,0.258)	0.446(0.088,0.798)	-0.483(-0.92,-0.046)	0.036(-0.26,0.338)	-0.026(-0.394,0.342)	-0.052(-0.36,0.247)	0.144(-0.313,0.592)	0.185(-0.105,0.481)	0.011(-0.271,0.291)	-0.425(-0.632,-0.216)	-0.025(-0.306,0.244)	0.081(-0.256,0.381)	0.02(-0.607,0.629)	-0.037(-0.438,0.376)
Asenapine	0.07(-0.258,0.377)	NA	0.517(0.126,0.891)	-0.413(-0.859,0.044)	0.106(-0.205,0.416)	0.044(-0.359,0.437)	0.018(-0.314,0.338)	0.215(-0.249,0.648)	0.255(-0.049,0.556)	0.082(-0.242,0.403)	-0.354(-0.59,-0.129)	0.045(-0.276,0.354)	0.151(-0.18,0.472)	0.09(-0.536,0.696)	0.033(-0.408,0.446)
Clozapine	-0.446(-0.798,-0.088)	-0.517(-0.891,-0.126)	NA	-0.93(-1.397,-0.459)	-0.41(-0.756,-0.059)	-0.472(-0.899,-0.032)	-0.499(-0.86,-0.127)	-0.302(-0.799,0.19)	-0.262(-0.603,0.084)	-0.435(-0.793,-0.06)	-0.871(-1.165,-0.583)	-0.471(-0.837,-0.109)	-0.366(-0.753,0)	-0.427(-1.059,0.234)	-0.484(-0.938,-0.016)
Fluphenazine	0.483(0.046,0.92)	0.413(-0.044,0.859)	0.93(0.459,1.397)	NA	0.519(0.1,0.918)	0.457(-0.036,0.941)	0.431(-0.009,0.87)	0.628(0.055,1.163)	0.668(0.222,1.087)	0.495(0.038,0.959)	0.058(-0.318,0.44)	0.458(0.018,0.9)	0.564(0.116,1.023)	0.503(-0.19,1.161)	0.446(-0.075,0.979)
Haloperidol	-0.036(-0.338,0.26)	-0.106(-0.416,0.205)	0.41(0.059,0.756)	-0.519(-0.918,-0.1)	NA	-0.062(-0.409,0.281)	-0.088(-0.397,0.233)	0.108(-0.348,0.554)	0.149(-0.132,0.437)	-0.025(-0.341,0.28)	-0.461(-0.667,-0.254)	-0.061(-0.374,0.236)	0.045(-0.273,0.35)	-0.016(-0.626,0.576)	-0.073(-0.477,0.341)
Loxapine	0.026(-0.342,0.394)	-0.044(-0.437,0.359)	0.472(0.032,0.899)	-0.457(-0.941,0.036)	0.062(-0.281,0.409)	NA	-0.026(-0.424,0.371)	0.17(-0.338,0.683)	0.211(-0.148,0.577)	0.037(-0.343,0.425)	-0.399(-0.709,-0.092)	0.001(-0.377,0.384)	0.107(-0.294,0.504)	0.046(-0.45,0.546)	-0.011(-0.489,0.482)
Lurasidone	0.052(-0.247,0.36)	-0.018(-0.338,0.314)	0.499(0.127,0.86)	-0.431(-0.87,0.009)	0.088(-0.233,0.397)	0.026(-0.371,0.424)	NA	0.196(-0.275,0.634)	0.237(-0.077,0.535)	0.063(-0.27,0.387)	-0.373(-0.606,-0.146)	0.027(-0.285,0.353)	0.133(-0.178,0.45)	0.072(-0.586,0.679)	0.015(-0.39,0.432)
Molindone	-0.144(-0.592,0.313)	-0.215(-0.648,0.249)	0.302(-0.19,0.799)	-0.628(-1.163,-0.055)	-0.108(-0.554,0.348)	-0.17(-0.683,0.338)	-0.196(-0.634,0.275)	NA	0.04(-0.361,0.455)	-0.133(-0.583,0.326)	-0.569(-0.963,-0.183)	-0.169(-0.623,0.276)	-0.064(-0.48,0.375)	-0.125(-0.802,0.572)	-0.182(-0.711,0.333)
Olanzapine	-0.185(-0.481,0.105)	-0.255(-0.556,0.049)	0.262(-0.084,0.603)	-0.668(-1.087,-0.222)	-0.149(-0.437,0.132)	-0.211(-0.577,0.148)	-0.237(-0.535,0.077)	-0.04(-0.455,0.361)	NA	-0.174(-0.474,0.115)	-0.61(-0.811,-0.418)	-0.21(-0.51,0.096)	-0.104(-0.401,0.181)	-0.165(-0.756,0.451)	-0.222(-0.626,0.182)
Paliperidone	-0.011(-0.291,0.271)	-0.082(-0.403,0.242)	0.435(0.06,0.793)	-0.495(-0.959,-0.038)	0.025(-0.28,0.341)	-0.037(-0.425,0.343)	-0.063(-0.387,0.27)	0.133(-0.326,0.583)	0.174(-0.115,0.474)	NA	-0.436(-0.665,-0.204)	-0.036(-0.362,0.283)	0.07(-0.249,0.403)	0.008(-0.621,0.633)	-0.049(-0.479,0.376)
Placebo	0.425(0.216,0.632)	0.354(0.129,0.59)	0.871(0.583,1.165)	-0.058(-0.44,0.318)	0.461(0.254,0.667)	0.399(0.092,0.709)	0.373(0.146,0.606)	0.569(0.183,0.963)	0.61(0.418,0.811)	0.436(0.204,0.665)	NA	0.4(0.168,0.621)	0.506(0.277,0.738)	0.444(-0.125,1.002)	0.387(0.044,0.735)
Quetiapine	0.025(-0.244,0.306)	-0.045(-0.354,0.276)	0.471(0.109,0.837)	-0.458(-0.9,-0.018)	0.061(-0.236,0.374)	-0.001(-0.384,0.377)	-0.027(-0.353,0.285)	0.169(-0.276,0.623)	0.21(-0.096,0.51)	0.036(-0.283,0.362)	-0.4(-0.621,-0.168)	NA	0.106(-0.197,0.412)	0.045(-0.583,0.647)	-0.012(-0.426,0.402)
Risperidone	-0.081(-0.381,0.256)	-0.151(-0.472,0.18)	0.366(0,0.753)	-0.564(-1.023,-0.116)	-0.045(-0.35,0.273)	-0.107(-0.504,0.294)	-0.133(-0.45,0.178)	0.064(-0.375,0.48)	0.104(-0.181,0.401)	-0.07(-0.403,0.249)	-0.506(-0.738,-0.277)	-0.106(-0.412,0.197)	NA	-0.061(-0.668,0.57)	-0.118(-0.539,0.297)
Trifluoperazine	-0.02(-0.629,0.607)	-0.09(-0.696,0.536)	0.427(-0.234,1.059)	-0.503(-1.161,0.19)	0.016(-0.576,0.626)	-0.046(-0.546,0.45)	-0.072(-0.679,0.586)	0.125(-0.572,0.802)	0.165(-0.451,0.756)	-0.008(-0.633,0.621)	-0.444(-1.002,0.125)	-0.045(-0.647,0.583)	0.061(-0.57,0.668)	NA	-0.057(-0.721,0.626)
Ziprasidone	0.037(-0.376,0.438)	-0.033(-0.446,0.408)	0.484(0.016,0.938)	-0.446(-0.979,0.075)	0.073(-0.341,0.477)	0.011(-0.482,0.489)	-0.015(-0.432,0.39)	0.182(-0.333,0.711)	0.222(-0.182,0.626)	0.049(-0.376,0.479)	-0.387(-0.735,-0.044)	0.012(-0.402,0.426)	0.118(-0.297,0.539)	0.057(-0.626,0.721)	NA

Appendix Table 11: League table for all comparisons in CA network. Informative priors from GP studies using a data based approach for β and moderate downweight for GP studies with high RoB.

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.065(-0.314,0.182)	0.457(0.122,0.767)	-0.536(-0.927,-0.128)	0.047(-0.216,0.312)	-0.018(-0.358,0.311)	-0.026(-0.271,0.224)	0.195(-0.186,0.57)	0.204(-0.03,0.45)	-0.002(-0.216,0.213)	-0.423(-0.59,-0.262)	-0.027(-0.262,0.215)	0.125(-0.13,0.376)	0.023(-0.522,0.572)	-0.141(-0.447,0.171)
Asenapine	0.065(-0.182,0.314)	NA	0.522(0.171,0.888)	-0.47(-0.88,-0.048)	0.113(-0.162,0.383)	0.047(-0.302,0.383)	0.039(-0.231,0.304)	0.26(-0.119,0.657)	0.269(-0.002,0.53)	0.063(-0.217,0.336)	-0.357(-0.547,-0.154)	0.038(-0.227,0.31)	0.19(-0.099,0.465)	0.089(-0.455,0.628)	-0.076(-0.392,0.267)
Clozapine	-0.457(-0.767,-0.122)	-0.522(-0.888,-0.171)	NA	-0.993(-1.467,-0.515)	-0.41(-0.754,-0.085)	-0.475(-0.893,-0.063)	-0.483(-0.828,-0.151)	-0.262(-0.697,0.156)	-0.253(-0.58,0.072)	-0.459(-0.809,-0.11)	-0.88(-1.164,-0.598)	-0.484(-0.826,-0.138)	-0.332(-0.686,0.011)	-0.434(-1.052,0.179)	-0.598(-0.986,-0.193)
Fluphenazine	0.536(0.128,0.927)	0.47(0.048,0.88)	0.993(0.515,1.467)	NA	0.583(0.189,0.961)	0.517(0.06,0.963)	0.509(0.063,0.926)	0.73(0.207,1.262)	0.739(0.306,1.14)	0.533(0.108,0.948)	0.113(-0.258,0.472)	0.508(0.092,0.919)	0.661(0.237,1.07)	0.559(-0.07,1.147)	0.395(-0.042,0.851)
Haloperidol	-0.047(-0.312,0.216)	-0.113(-0.383,0.162)	0.41(0.085,0.754)	-0.583(-0.961,-0.189)	NA	-0.065(-0.394,0.275)	-0.074(-0.351,0.201)	0.147(-0.25,0.581)	0.156(-0.095,0.417)	-0.05(-0.329,0.252)	-0.47(-0.676,-0.27)	-0.075(-0.348,0.204)	0.078(-0.202,0.352)	-0.024(-0.561,0.521)	-0.188(-0.525,0.165)
Loxapine	0.018(-0.311,0.358)	-0.047(-0.383,0.302)	0.475(0.063,0.893)	-0.517(-0.963,-0.06)	0.065(-0.275,0.394)	NA	-0.008(-0.373,0.338)	0.213(-0.223,0.66)	0.222(-0.124,0.57)	0.016(-0.329,0.371)	-0.405(-0.703,-0.113)	-0.009(-0.365,0.369)	0.143(-0.219,0.501)	0.041(-0.358,0.455)	-0.123(-0.524,0.277)
Lurasidone	0.026(-0.224,0.271)	-0.039(-0.304,0.231)	0.483(0.151,0.828)	-0.509(-0.926,-0.063)	0.074(-0.201,0.351)	0.008(-0.338,0.373)	NA	0.221(-0.168,0.579)	0.23(-0.035,0.497)	0.024(-0.244,0.303)	-0.396(-0.588,-0.201)	-0.001(-0.273,0.278)	0.151(-0.123,0.422)	0.05(-0.506,0.595)	-0.115(-0.429,0.236)
Molindone	-0.195(-0.57,0.186)	-0.26(-0.657,0.119)	0.262(-0.156,0.697)	-0.73(-1.262,-0.207)	-0.147(-0.581,0.251)	-0.213(-0.66,0.223)	-0.221(-0.579,0.168)	NA	0.009(-0.323,0.366)	-0.197(-0.586,0.204)	-0.617(-0.951,-0.283)	-0.222(-0.621,0.177)	-0.069(-0.412,0.282)	-0.171(-0.789,0.441)	-0.336(-0.753,0.106)
Olanzapine	-0.204(-0.45,0.03)	-0.269(-0.53,0.002)	0.253(-0.072,0.58)	-0.739(-1.14,-0.306)	-0.156(-0.417,0.095)	-0.222(-0.57,0.124)	-0.23(-0.497,0.035)	-0.009(-0.366,0.323)	NA	-0.206(-0.496,0.061)	-0.626(-0.809,-0.44)	-0.231(-0.494,0.046)	-0.079(-0.313,0.152)	-0.18(-0.729,0.362)	-0.345(-0.664,-0.009)
Paliperidone	0.002(-0.213,0.216)	-0.063(-0.336,0.217)	0.459(0.11,0.809)	-0.533(-0.948,-0.108)	0.05(-0.252,0.329)	-0.016(-0.371,0.329)	-0.024(-0.303,0.244)	0.197(-0.204,0.586)	0.206(-0.061,0.496)	NA	-0.42(-0.614,-0.236)	-0.025(-0.311,0.239)	0.127(-0.171,0.386)	0.026(-0.536,0.58)	-0.139(-0.455,0.197)
Placebo	0.423(0.262,0.59)	0.357(0.154,0.547)	0.88(0.598,1.164)	-0.113(-0.472,0.258)	0.47(0.27,0.676)	0.405(0.113,0.703)	0.396(0.201,0.588)	0.617(0.283,0.951)	0.626(0.44,0.809)	0.42(0.236,0.614)	NA	0.395(0.191,0.6)	0.548(0.343,0.737)	0.446(-0.082,0.976)	0.282(0.03,0.567)
Quetiapine	0.027(-0.215,0.262)	-0.038(-0.31,0.227)	0.484(0.138,0.826)	-0.508(-0.919,-0.092)	0.075(-0.204,0.348)	0.009(-0.369,0.365)	0.001(-0.278,0.273)	0.222(-0.177,0.621)	0.231(-0.046,0.494)	0.025(-0.239,0.311)	-0.395(-0.6,-0.191)	NA	0.152(-0.124,0.424)	0.051(-0.509,0.62)	-0.114(-0.425,0.229)
Risperidone	-0.125(-0.376,0.13)	-0.19(-0.465,0.099)	0.332(-0.011,0.686)	-0.661(-1.07,-0.237)	-0.078(-0.352,0.202)	-0.143(-0.501,0.219)	-0.151(-0.422,0.123)	0.069(-0.282,0.412)	0.079(-0.152,0.313)	-0.127(-0.386,0.171)	-0.548(-0.737,-0.343)	-0.152(-0.424,0.124)	NA	-0.102(-0.651,0.483)	-0.266(-0.592,0.085)
Trifluoperazine	-0.023(-0.572,0.522)	-0.089(-0.628,0.455)	0.434(-0.179,1.052)	-0.559(-1.147,0.07)	0.024(-0.521,0.561)	-0.041(-0.455,0.358)	-0.05(-0.595,0.506)	0.171(-0.441,0.789)	0.18(-0.362,0.729)	-0.026(-0.58,0.536)	-0.446(-0.976,0.082)	-0.051(-0.62,0.509)	0.102(-0.483,0.651)	NA	-0.164(-0.752,0.433)
Ziprasidone	0.141(-0.171,0.447)	0.076(-0.267,0.392)	0.598(0.193,0.986)	-0.395(-0.851,0.042)	0.188(-0.165,0.525)	0.123(-0.277,0.524)	0.115(-0.236,0.429)	0.336(-0.106,0.753)	0.345(0.009,0.664)	0.139(-0.197,0.455)	-0.282(-0.567,-0.03)	0.114(-0.229,0.425)	0.266(-0.085,0.592)	0.164(-0.433,0.752)	NA

Appendix Table 12: League table for all comparisons in CA network. Informative priors from GP studies using a data based approach for β and moderate downweight for GP studies evaluating interventions in $T_a - T_c$.

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.066(-0.325,0.189)	0.447(0.149,0.766)	-0.503(-0.895,-0.12)	0.037(-0.208,0.28)	-0.023(-0.355,0.302)	-0.078(-0.333,0.169)	0.113(-0.312,0.528)	0.166(-0.077,0.396)	0.01(-0.217,0.246)	-0.439(-0.613,-0.269)	-0.016(-0.271,0.238)	0.045(-0.224,0.318)	0.021(-0.567,0.628)	-0.035(-0.411,0.356)
Asenapine	0.066(-0.188,0.325)	NA	0.513(0.203,0.827)	-0.437(-0.844,-0.019)	0.103(-0.146,0.352)	0.043(-0.293,0.379)	-0.013(-0.277,0.249)	0.178(-0.269,0.596)	0.232(-0.006,0.485)	0.075(-0.188,0.35)	-0.373(-0.564,-0.183)	0.049(-0.229,0.334)	0.111(-0.167,0.385)	0.087(-0.507,0.69)	0.03(-0.346,0.414)
Clozapine	-0.447(-0.766,-0.149)	-0.513(-0.827,-0.203)	NA	-0.95(-1.398,-0.492)	-0.41(-0.707,-0.103)	-0.47(-0.857,-0.107)	-0.526(-0.849,-0.216)	-0.334(-0.8,0.101)	-0.281(-0.568,0.012)	-0.438(-0.753,-0.122)	-0.886(-1.14,-0.632)	-0.463(-0.783,-0.14)	-0.402(-0.726,-0.084)	-0.426(-1.053,0.19)	-0.482(-0.895,-0.03)
Fluphenazine	0.503(0.12,0.895)	0.437(0.019,0.844)	0.95(0.492,1.398)	NA	0.54(0.147,0.929)	0.48(0.021,0.936)	0.424(0.008,0.82)	0.616(0.09,1.113)	0.669(0.286,1.049)	0.513(0.108,0.916)	0.064(-0.297,0.428)	0.487(0.078,0.885)	0.548(0.123,0.959)	0.524(-0.162,1.195)	0.468(-0.016,0.965)
Haloperidol	-0.037(-0.28,0.208)	-0.103(-0.352,0.146)	0.41(0.103,0.707)	-0.54(-0.929,-0.147)	NA	-0.06(-0.386,0.264)	-0.116(-0.371,0.146)	0.076(-0.352,0.48)	0.129(-0.091,0.357)	-0.027(-0.263,0.221)	-0.476(-0.644,-0.31)	-0.053(-0.319,0.205)	0.008(-0.258,0.266)	-0.016(-0.617,0.551)	-0.072(-0.44,0.315)
Loxapine	0.023(-0.302,0.355)	-0.043(-0.379,0.293)	0.47(0.107,0.857)	-0.48(-0.936,-0.021)	0.06(-0.264,0.386)	NA	-0.056(-0.409,0.286)	0.135(-0.364,0.586)	0.188(-0.128,0.522)	0.032(-0.3,0.383)	-0.416(-0.695,-0.13)	0.006(-0.351,0.367)	0.068(-0.283,0.423)	0.044(-0.439,0.569)	-0.013(-0.455,0.451)
Lurasidone	0.078(-0.169,0.333)	0.013(-0.249,0.277)	0.526(0.216,0.849)	-0.424(-0.82,-0.008)	0.116(-0.146,0.371)	0.056(-0.286,0.409)	NA	0.191(-0.248,0.599)	0.244(0.005,0.488)	0.088(-0.178,0.35)	-0.36(-0.535,-0.176)	0.062(-0.201,0.336)	0.123(-0.153,0.39)	0.1(-0.5,0.712)	0.043(-0.333,0.427)
Molindone	-0.113(-0.528,0.312)	-0.178(-0.596,0.269)	0.334(-0.101,0.8)	-0.616(-1.113,-0.09)	-0.076(-0.48,0.352)	-0.135(-0.586,0.364)	-0.191(-0.599,0.248)	NA	0.053(-0.35,0.459)	-0.103(-0.516,0.325)	-0.551(-0.924,-0.139)	-0.129(-0.562,0.324)	-0.068(-0.473,0.349)	-0.092(-0.746,0.646)	-0.148(-0.635,0.378)
Olanzapine	-0.166(-0.396,0.077)	-0.232(-0.485,0.006)	0.281(-0.012,0.568)	-0.669(-1.049,-0.286)	-0.129(-0.357,0.091)	-0.188(-0.522,0.128)	-0.244(-0.488,-0.005)	-0.053(-0.459,0.35)	NA	-0.156(-0.395,0.086)	-0.605(-0.766,-0.447)	-0.182(-0.44,0.074)	-0.121(-0.361,0.125)	-0.145(-0.737,0.443)	-0.201(-0.58,0.169)
Paliperidone	-0.01(-0.246,0.217)	-0.075(-0.35,0.188)	0.438(0.122,0.753)	-0.513(-0.916,-0.108)	0.027(-0.221,0.263)	-0.032(-0.383,0.3)	-0.088(-0.35,0.178)	0.103(-0.325,0.516)	0.156(-0.086,0.395)	NA	-0.448(-0.638,-0.263)	-0.026(-0.285,0.247)	0.035(-0.243,0.311)	0.012(-0.596,0.622)	-0.045(-0.403,0.359)
Placebo	0.439(0.269,0.613)	0.373(0.183,0.564)	0.886(0.632,1.14)	-0.064(-0.428,0.297)	0.476(0.31,0.644)	0.416(0.13,0.695)	0.36(0.176,0.535)	0.551(0.139,0.924)	0.605(0.447,0.766)	0.448(0.263,0.638)	NA	0.422(0.223,0.627)	0.484(0.274,0.687)	0.46(-0.113,1.028)	0.403(0.077,0.752)
Quetiapine	0.016(-0.238,0.271)	-0.049(-0.334,0.229)	0.463(0.14,0.783)	-0.487(-0.885,-0.078)	0.053(-0.205,0.319)	-0.006(-0.367,0.351)	-0.062(-0.336,0.201)	0.129(-0.324,0.562)	0.182(-0.074,0.44)	0.026(-0.247,0.285)	-0.422(-0.627,-0.223)	NA	0.061(-0.237,0.341)	0.037(-0.57,0.632)	-0.019(-0.383,0.365)
Risperidone	-0.045(-0.318,0.224)	-0.111(-0.385,0.167)	0.402(0.084,0.726)	-0.548(-0.959,-0.123)	-0.008(-0.266,0.258)	-0.068(-0.423,0.283)	-0.123(-0.39,0.153)	0.068(-0.349,0.473)	0.121(-0.125,0.361)	-0.035(-0.311,0.243)	-0.484(-0.687,-0.274)	-0.061(-0.341,0.237)	NA	-0.024(-0.626,0.592)	-0.08(-0.472,0.341)
Trifluoperazine	-0.021(-0.628,0.567)	-0.087(-0.69,0.507)	0.426(-0.19,1.053)	-0.524(-1.195,0.162)	0.016(-0.551,0.617)	-0.044(-0.569,0.439)	-0.1(-0.712,0.5)	0.092(-0.646,0.746)	0.145(-0.443,0.737)	-0.012(-0.622,0.596)	-0.46(-1.028,0.113)	-0.037(-0.632,0.57)	0.024(-0.592,0.626)	NA	-0.056(-0.69,0.64)
Ziprasidone	0.035(-0.356,0.411)	-0.03(-0.414,0.346)	0.482(0.03,0.895)	-0.468(-0.965,0.016)	0.072(-0.315,0.44)	0.013(-0.451,0.455)	-0.043(-0.427,0.333)	0.148(-0.378,0.635)	0.201(-0.169,0.58)	0.045(-0.359,0.403)	-0.403(-0.752,-0.077)	0.019(-0.365,0.383)	0.08(-0.341,0.472)	0.056(-0.64,0.69)	NA

Appendix Table 13: League table for all comparisons in CA network. Informative priors from GP studies using expert's opinion for β and no downweight for GP studies.

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.162(-0.475,0.128)	0.334(-0.017,0.682)	-0.799(-1.272,-0.347)	-0.132(-0.436,0.157)	-0.375(-0.739,-0.032)	-0.052(-0.356,0.224)	0.134(-0.249,0.493)	0.183(-0.094,0.463)	-0.013(-0.257,0.228)	-0.435(-0.617,-0.259)	-0.092(-0.342,0.156)	0.157(-0.123,0.432)	-0.382(-0.799,0.032)	-0.268(-0.563,0.023)
Asenapine	0.162(-0.128,0.475)	NA	0.496(0.107,0.871)	-0.637(-1.131,-0.133)	0.03(-0.306,0.37)	-0.213(-0.589,0.171)	0.11(-0.213,0.43)	0.296(-0.12,0.721)	0.344(0.041,0.662)	0.149(-0.159,0.49)	-0.274(-0.513,-0.024)	0.07(-0.237,0.389)	0.318(-0.015,0.62)	-0.22(-0.676,0.229)	-0.106(-0.43,0.219)
Clozapine	-0.334(-0.682,0.017)	-0.496(-0.871,-0.107)	NA	-1.133(-1.654,-0.623)	-0.466(-0.818,-0.094)	-0.71(-1.134,-0.283)	-0.386(-0.755,0)	-0.2(-0.628,0.255)	-0.152(-0.472,0.193)	-0.347(-0.709,0.015)	-0.77(-1.068,-0.469)	-0.426(-0.785,-0.057)	-0.178(-0.522,0.195)	-0.716(-1.194,-0.257)	-0.602(-0.996,-0.237)
Fluphenazine	0.799(0.347,1.272)	0.637(0.133,1.131)	1.133(0.623,1.654)	NA	0.667(0.226,1.098)	0.424(-0.105,0.952)	0.747(0.239,1.236)	0.933(0.391,1.477)	0.981(0.492,1.466)	0.786(0.314,1.264)	0.363(-0.054,0.785)	0.707(0.232,1.177)	0.955(0.474,1.433)	0.417(-0.157,0.994)	0.531(0.032,1.038)
Haloperidol	0.132(-0.157,0.436)	-0.03(-0.37,0.306)	0.466(0.094,0.818)	-0.667(-1.098,-0.226)	NA	-0.243(-0.599,0.134)	0.08(-0.235,0.414)	0.267(-0.161,0.676)	0.315(0.005,0.612)	0.119(-0.199,0.439)	-0.303(-0.541,-0.055)	0.04(-0.274,0.349)	0.289(-0.022,0.599)	-0.25(-0.671,0.164)	-0.136(-0.456,0.201)
Loxapine	0.375(0.032,0.739)	0.213(-0.171,0.589)	0.71(0.283,1.134)	-0.424(-0.952,0.105)	0.243(-0.134,0.599)	NA	0.324(-0.048,0.689)	0.51(0.023,0.968)	0.558(0.171,0.914)	0.362(-0.013,0.753)	-0.06(-0.38,0.256)	0.283(-0.08,0.649)	0.532(0.166,0.905)	-0.006(-0.354,0.33)	0.107(-0.266,0.508)
Lurasidone	0.052(-0.224,0.356)	-0.11(-0.43,0.213)	0.386(0,0.755)	-0.747(-1.236,-0.239)	-0.08(-0.414,0.235)	-0.324(-0.689,0.048)	NA	0.186(-0.221,0.597)	0.234(-0.053,0.537)	0.039(-0.283,0.362)	-0.384(-0.598,-0.152)	-0.04(-0.331,0.274)	0.208(-0.084,0.515)	-0.33(-0.76,0.106)	-0.216(-0.538,0.105)
Molindone	-0.134(-0.493,0.249)	-0.296(-0.721,0.12)	0.2(-0.255,0.628)	-0.933(-1.477,-0.391)	-0.267(-0.676,0.161)	-0.51(-0.968,-0.023)	-0.186(-0.597,0.221)	NA	0.048(-0.278,0.389)	-0.148(-0.533,0.261)	-0.57(-0.906,-0.224)	-0.226(-0.593,0.161)	0.022(-0.34,0.389)	-0.516(-1.04,-0.006)	-0.402(-0.809,0.003)
Olanzapine	-0.183(-0.463,0.094)	-0.344(-0.662,-0.041)	0.152(-0.193,0.472)	-0.981(-1.466,-0.492)	-0.315(-0.612,-0.005)	-0.558(-0.914,-0.171)	-0.234(-0.537,0.053)	-0.048(-0.389,0.278)	NA	-0.196(-0.498,0.105)	-0.618(-0.824,-0.414)	-0.274(-0.558,-0.004)	-0.026(-0.286,0.235)	-0.564(-0.992,-0.136)	-0.45(-0.77,-0.144)
Paliperidone	0.013(-0.228,0.257)	-0.149(-0.49,0.159)	0.347(-0.015,0.709)	-0.786(-1.264,-0.314)	-0.119(-0.439,0.199)	-0.362(-0.753,0.013)	-0.039(-0.362,0.283)	0.148(-0.261,0.533)	0.196(-0.105,0.498)	NA	-0.422(-0.634,-0.197)	-0.079(-0.387,0.219)	0.17(-0.141,0.457)	-0.369(-0.808,0.069)	-0.255(-0.572,0.062)
Placebo	0.435(0.259,0.617)	0.274(0.024,0.513)	0.77(0.469,1.068)	-0.363(-0.785,0.054)	0.303(0.055,0.541)	0.06(-0.256,0.38)	0.384(0.152,0.598)	0.57(0.224,0.906)	0.618(0.414,0.824)	0.422(0.197,0.634)	NA	0.344(0.138,0.542)	0.592(0.379,0.794)	0.054(-0.318,0.431)	0.168(-0.057,0.389)
Quetiapine	0.092(-0.156,0.342)	-0.07(-0.389,0.237)	0.426(0.057,0.785)	-0.707(-1.177,-0.232)	-0.04(-0.349,0.274)	-0.283(-0.649,0.08)	0.04(-0.274,0.331)	0.226(-0.161,0.593)	0.274(0.004,0.558)	0.079(-0.219,0.387)	-0.344(-0.542,-0.138)	NA	0.248(-0.027,0.51)	-0.29(-0.714,0.129)	-0.176(-0.481,0.128)
Risperidone	-0.157(-0.432,0.123)	-0.318(-0.62,0.015)	0.178(-0.195,0.522)	-0.955(-1.433,-0.474)	-0.289(-0.599,0.022)	-0.532(-0.905,-0.166)	-0.208(-0.515,0.084)	-0.022(-0.389,0.34)	0.026(-0.235,0.286)	-0.17(-0.457,0.141)	-0.592(-0.794,-0.379)	-0.248(-0.51,0.027)	NA	-0.538(-0.945,-0.108)	-0.424(-0.733,-0.117)
Trifluoperazine	0.382(-0.032,0.799)	0.22(-0.229,0.676)	0.716(0.257,1.194)	-0.417(-0.994,0.157)	0.25(-0.164,0.671)	0.006(-0.33,0.354)	0.33(-0.106,0.76)	0.516(0.006,1.04)	0.564(0.136,0.992)	0.369(-0.069,0.808)	-0.054(-0.431,0.318)	0.29(-0.129,0.714)	0.538(0.108,0.945)	NA	0.114(-0.313,0.56)
Ziprasidone	0.268(-0.023,0.563)	0.106(-0.219,0.43)	0.602(0.237,0.996)	-0.531(-1.038,-0.032)	0.136(-0.201,0.456)	-0.107(-0.508,0.266)	0.216(-0.105,0.538)	0.402(-0.003,0.809)	0.45(0.144,0.77)	0.255(-0.062,0.572)	-0.168(-0.389,0.057)	0.176(-0.128,0.481)	0.424(0.117,0.733)	-0.114(-0.56,0.313)	NA

Appendix Table 14: League table for all comparisons in CA network. Informative priors from GP studies using expert's opinion for β and moderate downweight for GP studies with high RoB

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.174(-0.497,0.131)	0.644(0.301,0.995)	-0.569(-1.027,-0.123)	-0.026(-0.32,0.277)	-0.313(-0.686,0.059)	-0.049(-0.352,0.227)	0.144(-0.254,0.529)	0.208(-0.058,0.475)	0.001(-0.226,0.245)	-0.422(-0.609,-0.238)	-0.114(-0.37,0.141)	0.155(-0.12,0.42)	-0.324(-0.736,0.094)	-0.262(-0.54,0.03)
Asenapine	0.174(-0.131,0.497)	NA	0.817(0.439,1.192)	-0.395(-0.889,0.112)	0.147(-0.193,0.494)	-0.139(-0.553,0.264)	0.125(-0.21,0.453)	0.318(-0.095,0.739)	0.382(0.076,0.712)	0.175(-0.144,0.518)	-0.248(-0.485,0.007)	0.06(-0.268,0.384)	0.328(0.019,0.634)	-0.15(-0.599,0.291)	-0.089(-0.423,0.257)
Clozapine	-0.644(-0.995,-0.301)	-0.817(-1.192,-0.439)	NA	-1.213(-1.721,-0.701)	-0.67(-1.025,-0.299)	-0.957(-1.38,-0.529)	-0.693(-1.074,-0.31)	-0.5(-0.93,-0.059)	-0.436(-0.772,-0.101)	-0.643(-1.006,-0.281)	-1.066(-1.356,-0.766)	-0.758(-1.109,-0.423)	-0.489(-0.836,-0.153)	-0.967(-1.431,-0.496)	-0.906(-1.281,-0.524)
Fluphenazine	0.569(0.123,1.027)	0.395(-0.112,0.889)	1.213(0.701,1.721)	NA	0.543(0.111,0.993)	0.256(-0.279,0.766)	0.52(0.035,1)	0.713(0.179,1.245)	0.777(0.336,1.238)	0.57(0.087,1.021)	0.147(-0.273,0.559)	0.455(-0.005,0.911)	0.724(0.277,1.204)	0.245(-0.31,0.764)	0.306(-0.152,0.791)
Haloperidol	0.026(-0.277,0.32)	-0.147(-0.494,0.193)	0.67(0.299,1.025)	-0.543(-0.993,-0.111)	NA	-0.287(-0.66,0.065)	-0.023(-0.375,0.297)	0.17(-0.235,0.595)	0.234(-0.06,0.517)	0.027(-0.305,0.346)	-0.396(-0.633,-0.159)	-0.088(-0.412,0.223)	0.181(-0.141,0.493)	-0.297(-0.714,0.099)	-0.236(-0.582,0.087)
Loxapine	0.313(-0.059,0.686)	0.139(-0.264,0.553)	0.957(0.529,1.38)	-0.256(-0.766,0.279)	0.287(-0.065,0.66)	NA	0.264(-0.153,0.649)	0.457(-0.015,0.906)	0.521(0.159,0.897)	0.314(-0.058,0.685)	-0.109(-0.412,0.205)	0.199(-0.174,0.587)	0.468(0.089,0.844)	-0.011(-0.368,0.325)	0.05(-0.344,0.441)
Lurasidone	0.049(-0.227,0.352)	-0.125(-0.453,0.21)	0.693(0.31,1.074)	-0.52(-1,-0.035)	0.023(-0.297,0.375)	-0.264(-0.649,0.153)	NA	0.193(-0.223,0.601)	0.257(-0.045,0.575)	0.05(-0.258,0.374)	-0.373(-0.584,-0.127)	-0.065(-0.384,0.245)	0.203(-0.095,0.51)	-0.275(-0.691,0.166)	-0.214(-0.522,0.123)
Molindone	-0.144(-0.529,0.254)	-0.318(-0.739,0.095)	0.5(0.059,0.93)	-0.713(-1.245,-0.179)	-0.17(-0.595,0.235)	-0.457(-0.906,0.015)	-0.193(-0.601,0.223)	NA	0.064(-0.258,0.415)	-0.143(-0.536,0.27)	-0.566(-0.903,-0.221)	-0.258(-0.634,0.141)	0.011(-0.337,0.382)	-0.468(-0.983,0.041)	-0.407(-0.797,0.027)
Olanzapine	-0.208(-0.475,0.058)	-0.382(-0.712,-0.076)	0.436(0.101,0.772)	-0.777(-1.238,-0.336)	-0.234(-0.517,0.06)	-0.521(-0.897,-0.159)	-0.257(-0.575,0.045)	-0.064(-0.415,0.258)	NA	-0.207(-0.5,0.096)	-0.63(-0.83,-0.429)	-0.322(-0.614,-0.031)	-0.053(-0.306,0.191)	-0.532(-0.963,-0.144)	-0.471(-0.772,-0.16)
Paliperidone	-0.001(-0.245,0.226)	-0.175(-0.518,0.144)	0.643(0.281,1.006)	-0.57(-1.021,-0.087)	-0.027(-0.346,0.305)	-0.314(-0.685,0.058)	-0.05(-0.374,0.258)	0.143(-0.27,0.536)	0.207(-0.096,0.5)	NA	-0.423(-0.639,-0.211)	-0.115(-0.409,0.17)	0.154(-0.154,0.435)	-0.325(-0.77,0.1)	-0.264(-0.565,0.039)
Placebo	0.422(0.238,0.609)	0.248(-0.007,0.485)	1.066(0.766,1.356)	-0.147(-0.559,0.273)	0.396(0.159,0.633)	0.109(-0.205,0.412)	0.373(0.127,0.584)	0.566(0.221,0.903)	0.63(0.429,0.83)	0.423(0.211,0.639)	NA	0.308(0.096,0.519)	0.577(0.363,0.78)	0.098(-0.281,0.444)	0.16(-0.069,0.385)
Quetiapine	0.114(-0.141,0.37)	-0.06(-0.384,0.268)	0.758(0.423,1.109)	-0.455(-0.911,0.005)	0.088(-0.223,0.412)	-0.199(-0.587,0.174)	0.065(-0.245,0.384)	0.258(-0.141,0.634)	0.322(0.031,0.614)	0.115(-0.17,0.409)	-0.308(-0.519,-0.096)	NA	0.269(-0.005,0.549)	-0.209(-0.635,0.214)	-0.148(-0.455,0.161)
Risperidone	-0.155(-0.42,0.12)	-0.328(-0.634,-0.019)	0.489(0.153,0.836)	-0.724(-1.204,-0.277)	-0.181(-0.493,0.141)	-0.468(-0.844,-0.089)	-0.203(-0.51,0.095)	-0.011(-0.382,0.337)	0.053(-0.191,0.306)	-0.154(-0.435,0.154)	-0.577(-0.78,-0.363)	-0.269(-0.549,0.005)	NA	-0.478(-0.897,-0.066)	-0.417(-0.72,-0.099)
Trifluoperazine	0.324(-0.094,0.736)	0.15(-0.291,0.599)	0.967(0.496,1.431)	-0.245(-0.764,0.31)	0.297(-0.099,0.714)	0.011(-0.325,0.368)	0.275(-0.166,0.691)	0.468(-0.041,0.983)	0.532(0.144,0.963)	0.325(-0.1,0.77)	-0.098(-0.444,0.281)	0.209(-0.214,0.635)	0.478(0.066,0.897)	NA	0.061(-0.363,0.484)
Ziprasidone	0.262(-0.03,0.54)	0.089(-0.257,0.423)	0.906(0.524,1.281)	-0.306(-0.791,0.152)	0.236(-0.087,0.582)	-0.05(-0.441,0.344)	0.214(-0.123,0.522)	0.407(-0.027,0.797)	0.471(0.16,0.772)	0.264(-0.039,0.565)	-0.16(-0.385,0.069)	0.148(-0.161,0.455)	0.417(0.099,0.72)	-0.061(-0.484,0.363)	NA

Appendix Table 15: League table for all comparisons in CA network. Informative priors from GP studies using expert's opinion for β and moderate downweight for GP studies evaluating interventions in $T_a - T_c$.

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.176(-0.476,0.104)	0.641(0.321,0.951)	-0.583(-1.02,-0.166)	-0.029(-0.297,0.256)	-0.266(-0.602,0.065)	-0.06(-0.335,0.199)	0.115(-0.266,0.489)	0.205(-0.049,0.464)	0.005(-0.226,0.238)	-0.432(-0.608,-0.256)	-0.135(-0.37,0.1)	0.148(-0.108,0.407)	-0.283(-0.66,0.08)	-0.264(-0.54,0.001)
Asenapine	0.176(-0.104,0.476)	NA	0.818(0.483,1.182)	-0.406(-0.873,0.06)	0.147(-0.166,0.474)	-0.089(-0.462,0.27)	0.116(-0.192,0.416)	0.292(-0.105,0.674)	0.382(0.104,0.671)	0.182(-0.105,0.503)	-0.255(-0.474,-0.029)	0.041(-0.253,0.341)	0.324(0.027,0.617)	-0.107(-0.515,0.301)	-0.088(-0.382,0.209)
Clozapine	-0.641(-0.951,-0.321)	-0.818(-1.182,-0.483)	NA	-1.224(-1.694,-0.737)	-0.671(-0.996,-0.342)	-0.907(-1.293,-0.506)	-0.701(-1.031,-0.357)	-0.526(-0.956,-0.129)	-0.436(-0.741,-0.123)	-0.636(-0.98,-0.276)	-1.073(-1.338,-0.804)	-0.776(-1.101,-0.444)	-0.494(-0.82,-0.157)	-0.924(-1.369,-0.492)	-0.906(-1.24,-0.574)
Fluphenazine	0.583(0.166,1.02)	0.406(-0.06,0.873)	1.224(0.737,1.694)	NA	0.553(0.114,0.975)	0.317(-0.18,0.789)	0.523(0.059,0.962)	0.698(0.199,1.221)	0.788(0.358,1.23)	0.588(0.157,1.041)	0.151(-0.232,0.553)	0.447(-0.021,0.899)	0.73(0.294,1.181)	0.3(-0.241,0.823)	0.318(-0.134,0.775)
Haloperidol	0.029(-0.256,0.297)	-0.147(-0.474,0.166)	0.671(0.342,0.996)	-0.553(-0.975,-0.114)	NA	-0.236(-0.572,0.093)	-0.031(-0.348,0.26)	0.145(-0.252,0.539)	0.235(-0.048,0.513)	0.035(-0.279,0.32)	-0.402(-0.637,-0.192)	-0.106(-0.418,0.175)	0.177(-0.12,0.454)	-0.253(-0.64,0.132)	-0.235(-0.544,0.063)
Loxapine	0.266(-0.065,0.602)	0.089(-0.27,0.462)	0.907(0.506,1.293)	-0.317(-0.789,0.18)	0.236(-0.093,0.572)	NA	0.205(-0.159,0.56)	0.381(-0.044,0.814)	0.471(0.139,0.819)	0.271(-0.092,0.63)	-0.166(-0.453,0.125)	0.13(-0.219,0.492)	0.413(0.065,0.753)	-0.017(-0.337,0.304)	0.001(-0.372,0.357)
Lurasidone	0.06(-0.199,0.335)	-0.116(-0.416,0.192)	0.701(0.357,1.031)	-0.523(-0.962,-0.059)	0.031(-0.26,0.348)	-0.205(-0.56,0.159)	NA	0.175(-0.194,0.568)	0.265(-0.003,0.553)	0.066(-0.238,0.366)	-0.372(-0.574,-0.162)	-0.075(-0.347,0.205)	0.208(-0.061,0.486)	-0.223(-0.617,0.175)	-0.204(-0.483,0.086)
Molindone	-0.115(-0.489,0.266)	-0.292(-0.674,0.105)	0.526(0.129,0.956)	-0.698(-1.221,-0.199)	-0.145(-0.539,0.252)	-0.381(-0.814,0.044)	-0.175(-0.568,0.194)	NA	0.09(-0.245,0.441)	-0.11(-0.497,0.289)	-0.547(-0.87,-0.212)	-0.25(-0.62,0.131)	0.033(-0.324,0.378)	-0.398(-0.849,0.065)	-0.379(-0.771,0.008)
Olanzapine	-0.205(-0.464,0.049)	-0.382(-0.671,-0.104)	0.436(0.123,0.741)	-0.788(-1.23,-0.358)	-0.235(-0.513,0.048)	-0.471(-0.819,-0.139)	-0.265(-0.553,0.003)	-0.09(-0.441,0.245)	NA	-0.2(-0.483,0.081)	-0.637(-0.827,-0.457)	-0.341(-0.6,-0.071)	-0.058(-0.302,0.176)	-0.488(-0.869,-0.111)	-0.47(-0.74,-0.204)
Paliperidone	-0.005(-0.238,0.226)	-0.182(-0.503,0.105)	0.636(0.276,0.98)	-0.588(-1.041,-0.157)	-0.035(-0.32,0.279)	-0.271(-0.63,0.092)	-0.066(-0.366,0.238)	0.11(-0.289,0.497)	0.2(-0.081,0.483)	NA	-0.437(-0.643,-0.231)	-0.141(-0.423,0.124)	0.142(-0.131,0.423)	-0.288(-0.676,0.103)	-0.27(-0.569,0.011)
Placebo	0.432(0.256,0.608)	0.255(0.029,0.474)	1.073(0.804,1.338)	-0.151(-0.553,0.232)	0.402(0.192,0.637)	0.166(-0.125,0.453)	0.372(0.162,0.574)	0.547(0.212,0.87)	0.637(0.457,0.827)	0.437(0.231,0.643)	NA	0.297(0.098,0.499)	0.58(0.395,0.774)	0.149(-0.188,0.485)	0.168(-0.045,0.377)
Quetiapine	0.135(-0.1,0.37)	-0.041(-0.341,0.253)	0.776(0.444,1.101)	-0.447(-0.899,0.021)	0.106(-0.175,0.418)	-0.13(-0.492,0.219)	0.075(-0.205,0.347)	0.25(-0.131,0.62)	0.341(0.071,0.6)	0.141(-0.124,0.423)	-0.297(-0.499,-0.098)	NA	0.283(0.027,0.544)	-0.148(-0.533,0.241)	-0.129(-0.432,0.158)
Risperidone	-0.148(-0.407,0.108)	-0.324(-0.617,-0.027)	0.494(0.157,0.82)	-0.73(-1.181,-0.294)	-0.177(-0.454,0.12)	-0.413(-0.753,-0.065)	-0.208(-0.486,0.061)	-0.033(-0.378,0.324)	0.058(-0.176,0.302)	-0.142(-0.423,0.131)	-0.58(-0.774,-0.395)	-0.283(-0.544,-0.027)	NA	-0.431(-0.803,-0.045)	-0.412(-0.693,-0.13)
Trifluoperazine	0.283(-0.08,0.66)	0.107(-0.301,0.515)	0.924(0.492,1.369)	-0.3(-0.823,0.241)	0.253(-0.132,0.64)	0.017(-0.304,0.337)	0.223(-0.175,0.617)	0.398(-0.065,0.849)	0.488(0.111,0.869)	0.288(-0.103,0.676)	-0.149(-0.485,0.188)	0.148(-0.241,0.535)	0.431(0.045,0.803)	NA	0.019(-0.377,0.421)
Ziprasidone	0.264(-0.001,0.54)	0.088(-0.209,0.382)	0.906(0.574,1.24)	-0.318(-0.775,0.134)	0.235(-0.063,0.544)	-0.001(-0.357,0.372)	0.204(-0.086,0.483)	0.379(-0.008,0.771)	0.47(0.204,0.74)	0.27(-0.011,0.569)	-0.168(-0.377,0.045)	0.129(-0.158,0.432)	0.412(0.13,0.693)	-0.019(-0.421,0.377)	NA

Appendix Table 16: League table for all comparisons in CA network using a NMA model with non-informative priors.

	Aripiprazole	Asenapine	Clozapine	Fluphenazine	Haloperidol	Loxapine	Lurasidone	Molindone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Trifluoperazine	Ziprasidone
Aripiprazole	NA	-0.046(-0.517,0.444)	0.628(-0.142,1.404)	-1.555(-2.581,-0.524)	-0.193(-0.927,0.525)	-0.295(-1.284,0.633)	0.05(-0.388,0.512)	0.361(-0.212,0.97)	0.317(-0.149,0.769)	-0.036(-0.342,0.282)	-0.433(-0.718,-0.137)	-0.044(-0.381,0.293)	0.24(-0.2,0.66)	-0.259(-1.383,0.801)	-0.281(-0.744,0.224)
Asenapine	0.046(-0.44,0.517)	NA	0.674(-0.084,1.466)	-1.509(-2.529,-0.474)	-0.147(-0.916,0.621)	-0.249(-1.254,0.7)	0.096(-0.434,0.608)	0.407(-0.234,1.044)	0.362(-0.16,0.863)	0.01(-0.477,0.519)	-0.387(-0.762,-0.011)	0.001(-0.471,0.515)	0.286(-0.262,0.763)	-0.213(-1.321,0.908)	-0.235(-0.756,0.294)
Clozapine	-0.628(-1.404,0.142)	-0.674(-1.466,0.084)	NA	-2.183(-3.109,-1.171)	-0.821(-1.519,-0.16)	-0.923(-1.823,-0.014)	-0.578(-1.352,0.227)	-0.267(-1.009,0.509)	-0.312(-0.917,0.308)	-0.664(-1.439,0.154)	-1.061(-1.731,-0.351)	-0.673(-1.415,0.09)	-0.388(-1.072,0.31)	-0.887(-1.869,0.091)	-0.909(-1.721,-0.094)
Fluphenazine	1.555(0.524,2.581)	1.509(0.474,2.529)	2.183(1.171,3.109)	NA	1.362(0.642,2.089)	1.26(0.329,2.21)	1.604(0.555,2.706)	1.916(0.877,2.953)	1.871(0.958,2.833)	1.518(0.502,2.58)	1.122(0.12,2.152)	1.51(0.51,2.561)	1.795(0.828,2.733)	1.296(0.265,2.343)	1.274(0.223,2.373)
Haloperidol	0.193(-0.525,0.927)	0.147(-0.621,0.916)	0.821(0.16,1.519)	-1.362(-2.089,-0.642)	NA	-0.102(-0.718,0.535)	0.243(-0.507,1.026)	0.554(-0.186,1.301)	0.509(-0.046,1.115)	0.157(-0.605,0.938)	-0.24(-0.892,0.442)	0.148(-0.577,0.902)	0.433(-0.191,1.078)	-0.066(-0.812,0.705)	-0.088(-0.863,0.684)
Loxapine	0.295(-0.633,1.284)	0.249(-0.7,1.254)	0.923(0.014,1.823)	-1.26(-2.21,-0.329)	0.102(-0.535,0.718)	NA	0.345(-0.618,1.372)	0.656(-0.285,1.616)	0.611(-0.203,1.455)	0.259(-0.702,1.243)	-0.138(-1.033,0.834)	0.251(-0.684,1.232)	0.535(-0.337,1.448)	0.036(-0.413,0.494)	0.014(-0.953,1.018)
Lurasidone	-0.05(-0.512,0.388)	-0.096(-0.608,0.434)	0.578(-0.227,1.352)	-1.604(-2.706,-0.555)	-0.243(-1.026,0.507)	-0.345(-1.372,0.618)	NA	0.311(-0.324,0.948)	0.267(-0.28,0.78)	-0.086(-0.579,0.426)	-0.482(-0.85,-0.112)	-0.094(-0.58,0.393)	0.19(-0.358,0.677)	-0.309(-1.437,0.779)	-0.331(-0.852,0.186)
Molindone	-0.361(-0.97,0.212)	-0.407(-1.044,0.234)	0.267(-0.509,1.009)	-1.916(-2.953,-0.877)	-0.554(-1.301,0.186)	-0.656(-1.616,0.285)	-0.311(-0.948,0.324)	NA	-0.045(-0.509,0.41)	-0.397(-0.992,0.199)	-0.794(-1.316,-0.258)	-0.406(-0.985,0.196)	-0.121(-0.602,0.327)	-0.62(-1.655,0.444)	-0.642(-1.265,0.037)
Olanzapine	-0.317(-0.769,0.149)	-0.362(-0.863,0.16)	0.312(-0.308,0.917)	-1.871(-2.833,-0.958)	-0.509(-1.115,0.046)	-0.611(-1.455,0.203)	-0.267(-0.78,0.28)	0.045(-0.41,0.509)	NA	-0.353(-0.856,0.176)	-0.749(-1.119,-0.342)	-0.361(-0.819,0.119)	-0.076(-0.431,0.285)	-0.575(-1.516,0.392)	-0.597(-1.127,-0.04)
Paliperidone	0.036(-0.282,0.342)	-0.01(-0.519,0.477)	0.664(-0.154,1.439)	-1.518(-2.58,-0.502)	-0.157(-0.938,0.605)	-0.259(-1.243,0.702)	0.086(-0.426,0.579)	0.397(-0.199,0.992)	0.353(-0.176,0.856)	NA	-0.396(-0.757,-0.066)	-0.008(-0.461,0.406)	0.276(-0.213,0.738)	-0.223(-1.302,0.872)	-0.245(-0.764,0.289)
Placebo	0.433(0.137,0.718)	0.387(0.011,0.762)	1.061(0.351,1.731)	-1.122(-2.152,-0.12)	0.24(-0.442,0.892)	0.138(-0.834,1.033)	0.482(0.112,0.85)	0.794(0.258,1.316)	0.749(0.342,1.119)	0.396(0.066,0.757)	NA	0.388(0.08,0.705)	0.673(0.305,1.001)	0.174(-0.868,1.224)	0.152(-0.226,0.536)
Quetiapine	0.044(-0.293,0.381)	-0.001(-0.515,0.471)	0.673(-0.09,1.415)	-1.51(-2.561,-0.51)	-0.148(-0.902,0.577)	-0.251(-1.232,0.684)	0.094(-0.393,0.58)	0.406(-0.196,0.985)	0.361(-0.119,0.819)	0.008(-0.406,0.461)	-0.388(-0.705,-0.08)	NA	0.285(-0.162,0.696)	-0.214(-1.297,0.857)	-0.236(-0.718,0.288)
Risperidone	-0.24(-0.66,0.2)	-0.286(-0.763,0.262)	0.388(-0.31,1.072)	-1.795(-2.733,-0.828)	-0.433(-1.078,0.191)	-0.535(-1.448,0.337)	-0.19(-0.677,0.358)	0.121(-0.327,0.602)	0.076(-0.285,0.431)	-0.276(-0.738,0.213)	-0.673(-1.001,-0.305)	-0.285(-0.696,0.162)	NA	-0.499(-1.479,0.518)	-0.521(-1.011,0.009)
Trifluoperazine	0.259(-0.801,1.383)	0.213(-0.908,1.321)	0.887(-0.091,1.869)	-1.296(-2.343,-0.265)	0.066(-0.705,0.812)	-0.036(-0.494,0.413)	0.309(-0.779,1.437)	0.62(-0.444,1.655)	0.575(-0.392,1.516)	0.223(-0.872,1.302)	-0.174(-1.224,0.868)	0.214(-0.857,1.297)	0.499(-0.518,1.479)	NA	-0.022(-1.114,1.079)
Ziprasidone	0.281(-0.224,0.744)	0.235(-0.294,0.756)	0.909(0.094,1.721)	-1.274(-2.373,-0.223)	0.088(-0.684,0.863)	-0.014(-1.018,0.953)	0.331(-0.186,0.852)	0.642(-0.037,1.265)	0.597(0.04,1.127)	0.245(-0.289,0.764)	-0.152(-0.536,0.226)	0.236(-0.288,0.718)	0.521(-0.009,1.011)	0.022(-1.079,1.114)	NA