JANUS CHRISTIAN JAKOBSEN

DOCTORAL DISSERTATION

# SYSTEMATISING AND APPLYING ESSENTIAL METHODOLOGIES IN RANDOMISED CLINICAL TRIALS AND SYSTEMATIC REVIEWS WITHIN DIFFERENT MEDICAL SPECIALITIES

SDU
University of
Southern Denmark

CTU Copenhagen Trial Unit
Centre for Clinical Intervention Research

Denne afhandling er den 23. maj 2023 af Akademisk Råd,
Det Sundhedsvidenskabelige Fakultet, Syddansk Universitet
antaget til forsvar for den medicinske doktorgrad.

Ole Skøtt
h.a.dec.

Forsvaret finder sted den 6/11-2023 kl. 12.00 i lokale 0100,
Level 1, Building 34.1, SDU Odense, 5000 Odense C,
Danmark

Copenhagen Trial Unit, Centre for Clinical Intervention Research, Capital Region of Denmark, Rigshospitalet, Copenhagen University Hospital, Copenhagen, DK-2100, Denmark

Department of Regional Health Research, The Faculty of Health Sciences, University of Southern Denmark, Denmark

Illustrationer: Lupo Piva-Jakobsen & Anker Hermann-Schwenn

# ACKNOWLEDGEMENTS

I am deeply grateful to my family, especially my wife Rikke, my son Lupo, my stepson Anker, Krister Schwenn, my mother Lise, Simon Treschow, Kate Milvang, Mads Hermann, Rikke Opstrup Hansen and Dorrit Hermann for your continuous love and support.

I want to thank my mentor and primary collaborator, Christian Gluud. Meeting with him changed my life and was why I entered science. Christian has always supported me with great enthusiasm and inspiration. The Copenhagen Trial Unit has kindly provided means to enable the completion of this thesis, including an inspiring work environment with incredibly dedicated staff members.

I would like to thank all the people who work/ have worked at The Copenhagen Trial Unit and all my co-authors and collaborators during the last decade, especially Ole Mathiesen, Theis Lange, Niklas Jensen, Henrik Ancher Sørensen, Rikke Leth Larsen, Per Winkel, Peter Benzter, Jane Lindschou, Jørgen Hilden, Ronald Koretz, Gorm Greisen, Irving Kirsch, Lehana Thabane, Sebastian Simonsen, Silvio Garratini, Vittorio Bertele, Anders Kehlet, and Michael Hecht for excellent cooperation and hard work.

# TABLE OF CONTENTS

# SUMMARY

The crux of evidence-based medicine lies in randomised clinical trials, with systematic reviews of these trials regarded as the highest level of evidence assessing the effects of healthcare interventions.

There is a plethora of methodologies for analysing the results of systematic reviews and randomised clinical trials, with the ongoing development of new methods and novel applications of existing ones. However, certain methodologies are evidently more pivotal than others, presenting a challenge to researchers in selecting, combining, integrating, and applying individual methodologies. The present thesis considers how selected essential methodologies may be systematised in systematic reviews and randomised clinical trials within various medical specialities.

The first section of this thesis consists of eight theoretical papers aiming at systematising essential methodologies in systematic reviews and randomised clinical trials. Each topic of each theoretical paper has been carefully chosen based on experience in conducting trials and systematic reviews at The Copenhagen Trial Unit (www.ctu.dk) during the last two decades. This thesis pivots on the two articles presenting an overall guide on assessing if the statistical and clinical significance thresholds have been crossed in systematic reviews and randomised clinical trials. Six additional papers systematising other essential methodologies are also included in this thesis: 'When and how should multiple imputation be used for handling missing data in randomised clinical trials', 'Taking into account risks of random errors when analysing multiple outcomes in systematic reviews', 'Assessing assumptions for statistical analyses in randomised clinical trials',

‘Assessment of assumptions of statistical analysis methods in randomised clinical trials: the what, and how’, ‘Power estimations for non-primary outcomes in randomised clinical trials’, and ‘Count data analysis in randomised clinical trials’.

The second section of this thesis consists of three systematic reviews: ‘Direct-acting antivirals for chronic hepatitis C’, ‘Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder’, and ‘Drug-eluting stents versus bare-metal stents for acute coronary syndrome’. These three reviews are based on the methodologies described in the first section of this thesis. Two additional papers are included summarising the research findings of the systematic reviews to clinicians and patients, including recommendations on how patients should be treated.

This thesis explores the systemisation of essential methodologies in systematic reviews and randomised clinical trials across various medical fields. It will increase the validity of both randomised clinical trials and systematic reviews with meta-analysis if a systematised methodology is used.

# THIS THESIS IS BASED ON THE FOLLOWING PAPERS

**Jakobsen JC**, Wetterslev J, Winkel P, Lange T, Gluud C: Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. *BMC Med Res Methodol* 2014, 14:120.

**Jakobsen JC**, Gluud C, Winkel P, Lange T, Wetterslev J: The thresholds for statistical and clinical significance - a five-step procedure for evaluation of intervention effects in randomised clinical trials. *BMC Medical Research Methodology* 2014, 14(34).

**Jakobsen JC**, Gluud C, Wetterslev J, Winkel P: When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol* 2017, 17(1):162.

**Jakobsen JC**, Wetterslev J, Lange T, Gluud C: Viewpoint: taking into account risks of random errors when analysing multiple outcomes in systematic reviews [editorial]. *Cochrane Database of Syst Rev* 2016(3).

Nielsen EE., Norskov AK, Lange T, Thabane L, Wetterslev J, Beyersmann J, de Una-Alvarez J, Torri V, Billot L, Putter H, Winkel P, Gluud C, **Jakobsen JC**. Assessing assumptions for statistical analyses in randomised clinical trials. *BMJ Evid Based Med* 2019

Norskov AK, Lange T, Nielsen EE, Thabane L, Wetterslev J, Beyersmann J, de Una-Alvarez J, Torri V, Billot L, Putter H,

Winkel P, Gluud C, **Jakobsen JC**. Assessment of Assumptions of Statistical Analysis Methods in Randomised Clinical Trials: the What, and How. *BMJ Evid Based Med* 2019

**Jakobsen JC**, Ovesen C, Winkel P, Hilden J, Gluud C, Wetterslev J: Power estimations for non-primary outcomes in randomised clinical trials. *BMJ Open* 2019, 9(6):e027092.

**Jakobsen JC**, Tamborrino M, Winkel P, Haase N, Perner A, Wetterslev J, Gluud C: Count data analysis in randomised clinical trials. *J Biomet Biostat* 2015, 6(1):227.

**Jakobsen JC**, Nielsen EE, Feinberg J, Katakam KK, Fobian K, Hauser G, Poropat G, Djurisic S, Weiss KH, Bjelakovic M, Bjelakovic G, Klingenberg SL, Liu JP, Nikolova D, Koretz RL, Gluud C: Direct-acting antivirals for chronic hepatitis C. *Cochrane Database of Systematic Reviews* 2017 (9).

**Jakobsen JC**, Nielsen EE, Koretz RL, Gluud C: Do direct acting antivirals cure chronic hepatitis C? *BMJ* 2018, 361.

**Jakobsen JC**, Katakam KK, Schou A, Hellmuth SG, Stallknecht SE, Leth-Møller K, Iversen M, Banke MB, Petersen IJ, Klingenberg SL, Krogh J, Ebert SE, Timm A, Lindschou J, Gluud C: Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder. A systematic review with meta-analysis and Trial Sequential Analysis. *BMC Psychiatr* 2017, 17(1):58.

**Jakobsen JC**, Gluud C, Kirsch I: Should antidepressants be used for major depressive disorder? *BMJ Evid Based Med* 2019.

Feinberg J, Nielsen EE, Greenhalgh J, Hounsome J, Sethi NJ, Safi S, Gluud C, **Jakobsen JC**. Drug-eluting stents versus bare-metal stents for acute coronary syndrome. *Cochrane Database of Systematic Reviews* (8) 2017

## BACKGROUND

The crux of evidence-based medicine lies in randomised clinical trials [1], with systematic reviews of these trials regarded as the highest level of evidence assessing the effects of healthcare interventions [2]. There is a plethora of methodologies for analysing the results of systematic reviews and randomised clinical trials, with the ongoing development of new methods and novel applications of existing ones. However, certain methodologies are evidently more pivotal than others, presenting a challenge to researchers in selecting, combining, integrating, and applying individual methodologies. The present thesis considers how to systematise essential methodologies in systematic reviews and randomised clinical trials within different medical specialities.

# THRESHOLDS FOR STATISTICAL AND CLINICAL SIGNIFICANCE IN SYSTEMATIC REVIEWS OF RANDOMISED CLINICAL TRIALS

Systematic reviews synthesise the results from randomised clinical trials. Meta-analysis is the primary statistical method used in systematic reviews to analyse pooled results of trials [3, 4]. Some claim that systematic reviews should principally act as a catalyst for forming hypotheses and be used for planning upcoming randomised clinical trials [5-7]. Conversely, some regard systematic reviews with meta-analysis as the pinnacle of evidence for evaluating the impact of healthcare interventions [3, 4]. Research has demonstrated that meta-analyses of trials with low risks of bias offer more dependable results than single large trials [1, 8-12]. Inthout and colleagues assessed error rates for evaluations based on single, conventionally powered trials (80% or 90% power) against evaluations from random-effects meta-analyses of several smaller trials [10]. In scenarios where treatment was assumed to have no effect but heterogeneity was present, the error rates for a single trial escalated over ten times the standard rate [10]. On the other hand, the error rates for meta-analyses of multiple trials were accurate. If selective publication was prevalent, the error rates invariably escalated but remained generally lower for a series of trials compared to a single trial [10].

It seems evident that data from all randomised clinical trials conducted should be regarded as superior evidence compared to data from a single trial [1, 3, 13-16]. However, a systematic review with meta-analysis cannot be performed with the same scientific rigour as a randomised clinical trial with a pre-established high-quality methodology targeting a priori and quantitatively hypothesised intervention effect. Systematic review authors often know some eligible randomised clinical trials before formulating their protocol for the review, making the review methodology partially data-driven [17]. Nonetheless, recognising the inherent methodological constraints of a systematic review should lead to reducing these limitations and enhancing the remaining review methodology, which was the main goal of this article [17].

## What we showed

Approaches for evaluating the statistical and clinical significance of intervention effects in systematic reviews were considered. Consequently, an eight-step procedure was developed to ensure a more valid assessment of the outcomes of systematic reviews, striking a balance between simplicity and thoroughness [17]. Our process was rooted in and intended to augment The Cochrane Collaboration Handbook and the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) principles [13, 18-22]. This eight-step procedure may be incorporated into systematic review methodology or applied to verify the accuracy of results from previously published systematic reviews [23].

“To assess the statistical and clinical significance of results from systematic reviews, we proposed the following eight-step procedure [17]:

I. Calculate and report the confidence intervals and P-values from all fixed-effect and random-effects meta-analyses. The most conservative result should be the main result.

II. Explore the reasons behind substantial statistical heterogeneity by performing subgroup analyses and sensitivity analyses (see step 6).

III. Adjust the thresholds for significance (P-values and the confidence intervals from the meta-analyses and the risks of type I error in the trial sequential analysis) according to the number of primary outcomes.

IV. Calculate and report realistic diversity-adjusted required information sizes and analyse all primary and secondary outcomes in the review with Trial Sequential Analysis. Report if the trial sequential monitoring boundaries for benefit, harm, or futility are crossed [24, 25]. The Trial Sequential Analyses will adjust the confidence intervals and the thresholds for significance by relating the accrued data to the required information sizes [24, 25].

V. Calculate and report Bayes factor for the primary outcome/s based on a pre-specified anticipated intervention effect (same anticipated intervention effect as the one used to estimate the required information size (https://ctu.dk/tools-and-links/bayes-factor-calculation/). A Bayes factor less than 0.1 (a tenfold higher likelihood of compatibility with the alternative hypothesis than with the null hypothesis) may be used as threshold for significance.

VI. Use subgroup analysis and sensitivity analyses to assess the potential impact of systematic errors (bias).

VII. Assess the risk of publication bias (funnel plot).

VIII. Assess clinical significance of the review results if the prior seven steps have shown statistically significant results.“ [17].

We advocate that the prerogative and the incentive for conducting new trials and implementing new interventions in clinical practice should be a systematic review. The systematic review, as outlined, should rank highest in the hierarchy of evidence. However, conducting reviews without strict methodological rigour increases the risk of biased review results. To prevent hasty and incorrect conclusions, thorough statistical and clinical significance assessments in systematic reviews are necessary. Additionally, we outlined a systematic methodological procedure to enhance the validity and quality of the review results interpretation. The eight-step procedure has several advantages: (1) it encapsulates valid methodology concerning the specification and evaluation of significance thresholds in systematic reviews; (2) it systematically adjusts the significance thresholds based on the number of primary outcome comparisons and the portion of the required information size achieved; (3) it offers a likelihood ratio of the probability that a meta-analysis result aligns with the null hypothesis versus the probability that the result aligns with an anticipated intervention effect; (4) it curbs the tendency for review authors to both over- or underestimate the anticipated intervention effect (see step 4 and step 5); (5) it enables a more comprehensive assessment of the review results with a precise and detailed evaluation of imprecision, that could be used for a more accurate GRADE rating; (6) it obliges researchers and systematic review consumers to assess clinical significance.

The procedure has limitations. First, our eight-step procedure builds on existing, well-known techniques, but we need comprehensive comparative research that contrasts this process with standard practice. Furthermore, our recommended pragmatic method for multiplicity adjustment (which involves dividing 0.05 by the number mid-way between 1 and the number of primary outcomes comparisons) must have an evidence-based foundation. We hinge our strategy on the idea that the ‘correct’ threshold for multiplicity-adjusted significance falls somewhere between the unadjusted and Bonferroni-adjusted thresholds (as discussed in step 3). Yet, many systematic reviewers do not adjust the significance thresholds per outcome comparisons, a practice that seems inferior to both the conservative Bonferroni adjustment and our recommended pragmatic approach.

As outlined in step 3, it would be necessary to estimate correlations between the co-primary outcomes to make more accurate adjustments to the significance thresholds in systematic reviews (refer to step 3). Such correlations are often unknown, and incorrect assumptions about correlations could yield inaccurate results. Last, the necessary information size, Trial Sequential Analysis, and the size of the Bayes factor are heavily influenced by the anticipated intervention effect selection, which is often challenging to pre-determine. Sensitivity analyses using various anticipated intervention effect estimates are often necessary to mitigate this drawback. For instance, as an additional trial sequential analysis, the point estimate from the meta-analysis of previous trials, or the limit of the 95% confidence interval closest to no effect may be used as anticipated intervention effects; or a Bayes factor utilising a smaller, more sceptical, assumed intervention effect, such as a relative risk that is halfway between the anticipated intervention effect used to calculate the necessary information size and 1.0 [17]. The main limitation of our method, and of Bayesian analyses in general, is the uncertainty tied to quantifying the anticipated intervention effects. To optimally and objectively estimate the anticipated intervention effects, we suggest, among other options, basing the estimate on prior relevant randomised clinical trials. However, if this is done, the trials used to estimate the anticipated intervention effects will likely be reused in the full review analysis, meaning the estimation of the necessary information size becomes an 'adaptive' estimation [26, 27]. As a result, the risk of a type I error could rise [26, 27]. Because of this risk of circular logic, it is theoretically necessary to adjust the required information size by applying a penalty for the weight of data from prior systematic reviews (or trials) [26, 27]. If the assumed intervention effect estimate is based on estimating a 'minimal important difference, then no adjustments would be required. We recognise the theoretical need for such further adjustments, but this would complicate review analyses due to the varying methods used to quantify the anticipated intervention effects. Moreover, our suggestions are already significantly tightening the significance thresholds in systematic reviews. If these thresholds are too stringent, there is a risk of 'throwing the baby out with the bath water' [17].

Post-hoc modifications and incorrect assessments of the alternative hypothesis could impact the Trial Sequential Analysis, the necessary data size, and the Bayes factor, especially if these are not explicitly predefined in a protocol published before the systematic review's onset. The issues related to the use of anticipated intervention effects may be mitigated if the expected intervention outcomes are distinctly outlined in a review protocol that has been published and if predefined sensitivity analyses evaluate the uncertainty of the anticipated intervention effects' estimations. The unpredictability of estimating the anticipated intervention effects poses a significant challenge and limitation but is an unavoidable necessity. Without estimates of an anticipated intervention effect, it is impossible to compute the necessary information size and adjusted significance thresholds [17].

Adopting our eight-step procedure may result in fewer interventions that appear beneficial, potentially delaying the incorporation of effective interventions into clinical practice. We acknowledge these risks, but when weighed against the current practice of deploying strategies based on insufficient evidence, we believe a more cautious approach is ethically more defensible [11, 25, 28-31]. We maintain that the benefits of our procedure surpass its drawbacks. Healthcare researchers must provide compelling evidence of more benefit than harm before implementing interventions in clinical practice [17]. Adhering to this suggested eight-step procedure could enhance the reliability of intervention effect assessments in systematic reviews [17].

# THRESHOLDS FOR STATISTICAL AND CLINICAL SIGNIFICANCE IN RANDOMISED CLINICAL TRIALS

In randomised clinical trials, statistical analyses are typically conducted using the frequentist paradigm. Using this approach, a significant difference in effect is declared when a test statistic value surpasses a threshold, suggesting that it is unlikely that the trial results are produced by zero difference in effect between the compared interventions, or in other words, that the null hypothesis is accurate [32]. A P-value under 5% has been the standard threshold for statistical significance in clinical intervention research since Fisher warned against precisely that in 1955 [33-35]. Despite being easy to calculate, P-values are frequently misinterpreted [36, 37] and misapplied [30, 31, 38].

## What we showed?

Various methods were considered for evaluating the statistical and clinical significance of intervention effects in randomised clinical trials. To balance simplicity and comprehensiveness, a five-step procedure was developed [14].

“To assess the statistical significance and the clinical significance of results from randomised clinical superiority trials, we propose a five-step procedure:

I. Calculate and report the confidence intervals and the exact P- values for all pre-specified outcome comparisons. A P- value less than 0.05 may be chosen as threshold for statistical significance for the primary outcome only if 0.05 has been used as the acceptable risk of type I error in the sample size calculation and the sample size has been reached.

II. Calculate and report the Bayes factor for the primary outcome (or outcomes) based on the hypothesised intervention effect used in the sample size estimation. If the intervention effect hypothesised in the sample size calculation is not based on results from systematic reviews or randomised clinical trials, then calculate an additional sceptical Bayes factor using a smaller hypothesised intervention effect, e.g. a relative risk half-way between 1.0 and the intervention effect hypothesised in the sample size calculation. A Bayes factor less than 0.1, indicating a ten-fold higher likelihood of compatibility with the alternative hypothesis than the likelihood of compatibility with the null hypothesis, may be chosen as a threshold for supporting the alternative hypothesis.

III. If the a priori estimated sample size has not been reached or interim analyses have been performed, then adjust the confidence intervals and the P-values accordingly.

IV. If more than one outcome is used, if more than two intervention groups are compared, or if the primary outcome is assessed at multiple time points (and just one of these outcome comparisons must be significant to reject the overall null hypothesis), then the confidence intervals and the P-values should be adjusted accordingly.

V. Assess and report clinical significance of the results if all of the first four steps of the five-point procedure have shown statistical significance.“ [14]

Our five-step procedure comes with several strengths [14].

The five-step procedure is based on widely recognised methodologies. It provides a ratio of the probability that a trial result is compatible with the null hypothesis divided by the probability that the result is consistent with the intervention effect hypothesised in the sample size calculation. Our procedure considers risks of random errors, including issues of multiplicity, and compels investigators, and those who utilise clinical research, to assess clinical significance. A potential limitation of Bayesian statistical analyses is that it can be challenging to verify modelling assumptions, e.g., whether assumed distributions behind the statistical analyses are appropriate. It is a strength of our proposed procedure that if the assumptions behind the initial analysis methods (e.g. logistic regression or survival analysis) are fulfilled, then our five-step procedure can be legitimately applied without further testing.

Our five-step procedure also carries some limitations [14].

First, we have made suggestions for interpreting the outcomes of a single randomised clinical trial considering typically limited previous evidence. Studies have demonstrated that it is usually imprudent to base diagnostic, prognostic, preventive, or therapeutic interventions on data from one or a few trials [1, 31], and our procedure does not alter this fact in any way. Our goal is to present a simple assessment method to enhance the reliability of results from a single randomised clinical trial. Still, our procedure only provides a solution to some problems. Clinical decision-making should primarily depend on systematic reviews of all randomised clinical trials with minimal risks of bias, including meta-analyses, Trial Sequential Analyses, and reached consensus on clinical significance. Also, calculating the Bayes factor and evaluating clinical significance in a systematic review scenario could become critical.

Second, our suggested approach and our interpretation of the Bayes factor is simplified [14]. Alternatively, Bayes factors could be calculated based on the likelihood ratio of the trial outcome being consistent with the null hypothesis versus its compatibility with a set of plausible alternate hypotheses. A comprehensive full Bayesian analysis could also be employed to evaluate trial results, estimating the posterior odds of an alternate hypothesis being true against the null hypothesis, given the observed data and any prior information available. Employing comprehensive Bayesian statistics holds significant methodological benefits over traditional frequentist statistics, and the results from a thorough Bayesian analysis could occasionally present a low posterior probability for the alternate hypothesis, contradicting a low Bayes factor which may falsely suggest otherwise. However, incorporating Bayesian statistical analyses does increase the methodological complexity, potentially causing research findings to be sensitive to seemingly harmless assumptions, which may hinder taking potentially valid trial results into account. Essentially, incorporating Bayesian statistical analyses would necessitate a shift in methodological paradigm, including adopting intricate Bayesian statistical analysis plans and using specialised Bayesian statistical software like WinBUGS [14].

Third, it is necessary to define an alternative hypothesis to the null hypothesis when calculating the Bayes factor [14]. The definition of the alternative hypothesis often involves an element of subjectivity, a fact that deters many researchers from adopting a Bayesian standpoint. It has been proposed that the alternative hypothesis could be identified as 'uniformly most powerful Bayesian tests' where the alternative hypothesis is defined as an average value of any hypothetical intervention effect resulting in a Bayes factor below a given threshold [14]. This approach is attractive as it does not necessitate subjective assumptions about the alternative hypothesis. However, it presents an issue as it excludes potentially valuable data from previous randomised trials or systematic reviews in defining the alternative hypothesis [14]. Additionally, this method is essentially for single-parameter exponential family models and does not offer any methodological benefits over simply using the P-value as a significance threshold [14]. The proponent of the 'uniformly most powerful Bayesian tests' recommends utilising smaller P-value thresholds (0.005 or 0.001) to prevent false positive results, which appears to be a viable alternative to our computation and application of the Bayes factor. We have elected to use the intervention effect postulated in the sample size calculation as the alternate hypothesis, firmly linking the pre-planned trial design to interpreting the analysis results of the trial outcomes. Most trials already comprise a pre-established sample size calculation, including an anticipated intervention effect estimate. Therefore, new assumptions are not required to compute this Bayes factor. However, it is a distinct drawback that the Bayes factor can be affected by post hoc adjustments and incorrect quantifications of the alternate hypothesis [14].

Last, our procedure is grounded in an already well-established methodology. However, there is yet to be empirical evidence to evaluate the validity of the procedure [14].

We contend that the strengths of the procedure outweigh the limitations [14]. Healthcare researchers must provide solid evidence of more benefits than harms before interventions are implemented in clinical practice. If the proposed five-step procedure is adhered to, it may enhance the reliability of assessments of intervention effects in randomised clinical trials [14].

## WHEN AND HOW SHOULD MULTIPLE IMPUTATION BE USED FOR HANDLING MISSING DATA IN RANDOMISED CLINICAL TRIALS

The key strength of randomised clinical trials is that the random allocation of participants results in similar baseline characteristics in the compared groups – if enough participants are randomised [39, 40]. Consequently, the comparison groups in a sufficiently large randomised clinical trial are expected to be similar in all known and unknown prognostic characteristics at baseline [39, 40]. To keep this baseline similarity intact, the intention-to-treat principle should be employed in analysing randomised trials [39]. However, the validity of trial results may be jeopardised if certain participants are not included in the analysis, causing baseline differences between the groups in the analysis [39]. Inferences from randomised clinical trials may be significantly affected by missing data, especially if the missingness is not random and not dealt with correctly [41, 42]. The potential bias from missing data relies on why the data are missing and the analytical approach used to deal with the missing data. Thus, planning and attention are required when analysing trial data with missing data.

Three common mechanisms cause missing data: missing completely at random (MCAR); missing at random (MAR); and missing not at random (MNAR) [41-43]. The missing data mechanism may neither rely on the observed nor the missing data. In this case, it is said to be missing completely at random (MCAR) [42, 43]. MCAR increases standard errors due to smaller sample sizes but does not introduce bias [42]. In this scenario, the incomplete datasets represent the whole dataset [42]. More commonly, the missingness mechanism might depend on the observed data [42]. If it only relies on the observed data, the missing data are considered missing at random (MAR) based on the observed data [42]. MAR allows for predicting missing values from participants with complete data [42]. If the mechanism relies on the missing data, even considering the observed data, the data are categorised as missing not at random (MNAR) [42, 43]. The MAR and MNAR conditions cannot be differentiated based on the observed data as the missing data are, by definition, unknown, and it cannot be determined if the observed data can predict the unknown data [42, 43].

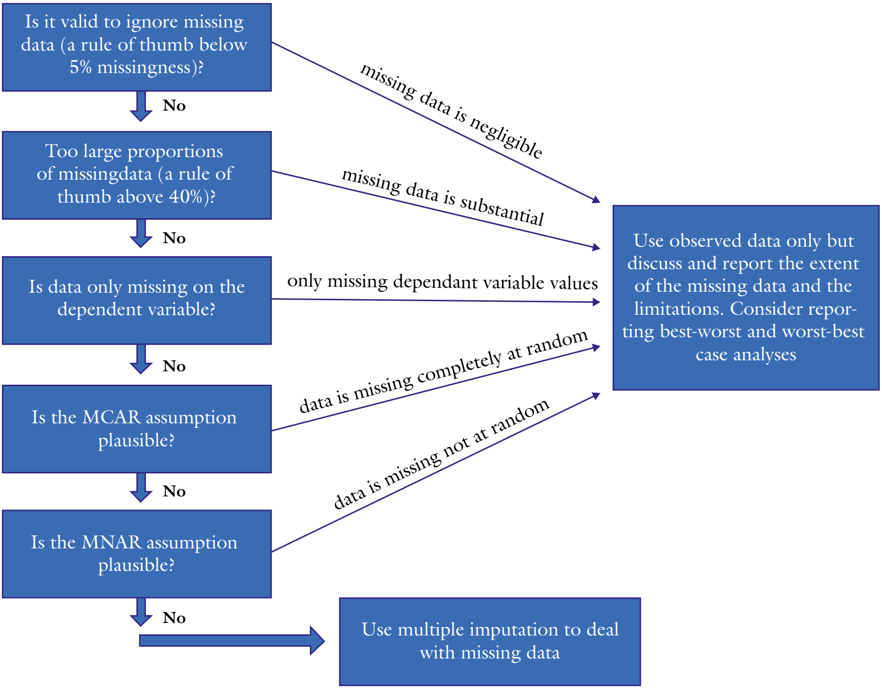
Methods like multiple imputations or full information direct maximum likelihood often result in unbiased outcomes when MAR is present. However, in some cases, the MAR assumption might not be clinically relevant [42]. As such, conducting sensitivity analyses is frequently required to evaluate the potential effect that MNAR might pose on the estimated outcomes [41, 44].

Dealing with missing data and employing multiple imputations is complex and a distinct topic [45]. We studied relevant previous studies based on searches of the literature [45]. We checked the reference lists of known studies for documents (theoretical articles, empirical studies, simulation studies, etc.) on how to deal with missing data when analysing randomised clinical trials. In addition, we searched PubMed (last search 14th September 2017), which resulted in 166 studies using the keywords 'missing data', 'randomi*', and 'statistical analysis'. After this, we developed the following flow charts [45].

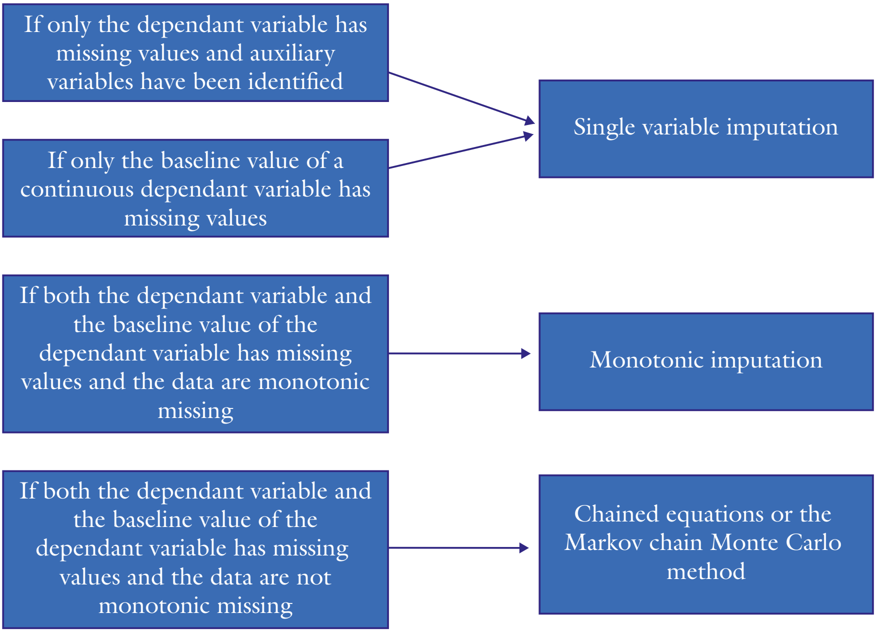
## What we showed

Based on group discussions, a review of included papers on this topic, and our personal experience in analysing results of randomised clinical trials, we have furnished a hands-on manual with flowcharts for handling missing data while analysing the results of randomised clinical trials [45].

Flowchart: when should multiple imputation be used [45]?


Is it valid to ignore missing data (a rule of thumb below 5% missingness)?
No
Too large proportions of missingdata (a rule of thumb above 40%)?
No
Is data only missing on the dependent variable?
No
Is the MCAR assumption plausible?
No
Is the MNAR assumption plausible?
No
missing data is negligible
missing data is substantial
only missing dependant variable values
data is missing completely at random
data is missing not at random
Use observed data only but discuss and report the extent of the missing data and the limitations. Consider repor-ting best-worst and worst-best case analyses
Use multiple imputation to deal with missing data

## Flowchart: which multiple imputation method should be used [45]?


If only the dependant variable has missing values and auxiliary variables have been identified
If only the baseline value of a continuous dependant variable has missing values
Single variable imputation
If both the dependant variable and the baseline value of the dependant variable has missing values and the data are monotonic missing
Monotonic imputation
If both the dependant variable and the baseline value of the dependant variable has missing values and the data are not monotonic missing
Chained equations or the Markov chain Monte Carlo method

Missing data can always pose a challenge when analysing the results of a trial. Even when the data are missing completely at random (MCAR), there can be a significant decrease in the statistical power [45]. It is crucial for trialists to thoroughly consider and discuss these potential shortcomings due to missing data. As always, prevention is better than cure. Hence, trials should be strategically focused and practical to ensure preventative measures. Interpreting trial outcomes based on incomplete data should always be done carefully. The inability to distinguish between missing at random (MAR) and missing not at random (MNAR) means that the reliability of the assumptions behind methods, like multiple imputation, may always be challenged. And in cases where data are MNAR, there are no established methods to handle the missing data appropriately. Nonetheless, the best-worst and worst-best case assessments always provide the most extensive range of uncertainty for dichotomised data and a probable uncertainty range for continuous data, considering 95% of the normally distributed observed data. The primary conclusion on the effects of interventions should reflect this demonstrated range of uncertainty. By systematically following our practical guide and flowcharts for when and how to use multiple imputation, the validity of managing missing data can be enhanced.

## TAKING INTO ACCOUNT RISKS OF RANDOM ERRORS WHEN ANALYSING MULTIPLE OUTCOMES IN SYSTEMATIC REVIEWS

Authors of reviews must be cautious not to focus excessively on the statistical significance of a meta-analysis result [46]. The clinical significance and positive or negative implication of an effect estimate should only be emphasised if the statistical assessment is conclusive beyond reasonable doubt [46]. Confidence intervals excluding 1.0 for a binary outcome result or 0.0 for a continuous outcome result, along with corresponding P values, are commonly used as markers of statistical significance [46]. The average number of outcome comparisons in Cochrane Reviews is 12 [47]. Given 12 independent hypotheses and a 5% significance level, the chance of making one or more type I errors exceeds 45% if all nulls are accurate [47]. It is essential to avoid wrongly rejecting the overall null hypothesis of a review to prevent incorrect confirmations of intervention effects caused by multiple outcome comparisons. In other words, the overall type I error should be capped at a maximum of 5% or less [17]. This implies employing one primary outcome evaluated with a traditional 95% confidence interval or at a P-value threshold of 5%, as long as the necessary information size has been met [17]. The overall likelihood of inaccurately rejecting the null hypothesis for at least one outcome in a review will rise if more than one outcome are evaluated or if an outcome is evaluated at different time points [17]. Consequently, if review authors can select and accentuate single results among multiple comparisons, there will likely be a heightened risk of erroneously presenting statistical evidence of an intervention effect [17]. We appreciate the Cochrane Handbook's suggestion of adopting up to three primary outcomes, such as all-cause mortality, health-related quality of life, and serious adverse events (https://training.cochrane.org/handbook). Most systematic reviews do indeed assess multiple primary outcomes or evaluations of outcomes at multiple time points. However, the thresholds for statistical significance (confidence intervals and P values) in meta-analyses are seldom correctly adjusted to align with the number of outcome comparisons [17]. Without the correct adjustments, claims of statistical significance could be incorrect [17].

## What we showed

Multiple primary outcomes bring about issues of multiplicity. Yet, these could be resolved if all primary outcomes were required to meet both clinical and statistical significance for the intervention to be deemed successful [17, 46]. Even though this is a conservative approach, it is feasible. A similar conservative method, the Bonferroni procedure, splits the pre-set P-value threshold, typically 0.05, by the number of outcomes being compared [17, 46]. For instance, with three primary outcomes, this would result in an adjusted P-value threshold of 0.016, equal to utilising a 98.4% confidence interval (calculated by deducting the adjusted P-value threshold from 1.00), excluding 1.0 for dichotomous outcomes or 0.0 for continuous outcomes as the significance threshold [17, 46]. The Bonferroni adjustment could be suitable if there is no or minimal correlation among the multiple primary outcomes. However, it may be overly conservative when outcomes are interconnected and correlated. Other legitimate threshold adjustment methods are available (e.g., Hommel's method, the fixed-sequence procedure, the fallback procedure, and Holm's procedure). However, they do not account for the correlations between outcomes and are likely too conservative. These methods enable the spread of a 5% risk of type I error across different outcomes but have shortcomings. Some meta-analytic null hypotheses will be accepted regardless of results, and very low secondary outcome significance thresholds must be applied [46]. Precise adjustments of statistical significance thresholds require a valid correlation estimate between the co-primary outcomes [46]. For example, estimating the correlation between all-cause mortality and serious adverse events would be necessary if they were chosen as co-primary outcomes. Several relatively complex methods that can consider the outcome correlations exist (e.g. bootstrapping and multivariate permutation methods). Although primarily developed for observational studies featuring thousands of outcomes and for analysing randomised clinical trials, these methods may apply to meta-analyses [46, 48]. However, correlations between different outcomes are often unknown, can differ between studies, and inaccurate or data-driven correlation estimates may yield erroneous results.

When adjusting for multiplicity, the 'correct' threshold for significance in meta-analysis findings will likely lie between the unadjusted threshold (usually 0.05) and the Bonferroni-adjusted threshold [17, 46]. Meta-analysis results may then be classified as: (1) 'statistically significant' ' if the P value falls below the Bonferroni-adjusted threshold (equivalent to a confidence interval of 1 minus the Bonferroni-adjusted P value threshold); (2) 'uncertain statistical significance' if the P value lies between the Bonferroni-adjusted threshold and 0.05; and (3) 'not statistically significant" if the P value is equal to or greater than 0.05. However, the downside of this method is that it cannot use adjusted P-value thresholds (used as type I error risks) to calculate a necessary information size beforehand at the protocol stage [17, 46]. We propose splitting the 0.05 probability by the value mid-way between 1 (no adjustment) and the number of primary outcome comparisons (the Bonferroni adjustment) [17]. This will yield a multiplicity-adjusted P-value threshold of 0.05 for one primary outcome (equating to a 95% confidence interval), 0.033 for two primary outcomes (equating to a 96.7%

confidence interval), and 0.025 for three primary outcomes (equating to a 97.5% confidence interval). Presenting confidence intervals appear to be a more suitable and comprehensible way to show statistical uncertainty. However, confidence intervals do not inherently offer more information than implicitly provided by the estimated effect and the P value [17, 49]. Assuming the necessary data are accessible, the confidence interval and observed effect size can be calculated from the P value, and vice versa [17, 49].

This pragmatic approach may also be applied independently to secondary outcomes and subgroup analyses [17]. For instance, if seven secondary outcomes are used, the significance thresholds for these outcomes could be 0.05/4 = 0.0125 (equating to a 98.75% confidence interval) [17].

Our pragmatic approach has clear limitations [17]. It fails to consider differing correlations among various results and may be too conservative or too loose, hinging on the correlations of outcomes [17]. Neither empirical research nor simulation studies form the basis of our suggestions. Nonetheless, in most scenarios with correlated results, our approach might both be better than no threshold adjustment and threshold adjustments assuming no correlation between outcomes [17]. Adjusting the statistical significance thresholds based on the number of outcome comparisons could encourage review authors to limit the selection of outcomes to only those most important to patients and choose primary results that will assist clinicians in determining the necessity of the intervention [17]. Selecting diverse sets of primary outcomes leads to responses to various clinical queries. Hence these factors must be meticulously considered [17]. Enhancing the selection of patient-important outcomes for decision-making might subsequently benefit both patients and the healthcare sector in general [17]. Some have proposed that systematic review outcomes should be viewed as hypothesis-generating and should only cater to the design of future randomised clinical trials [5-7]. To maintain the systematic review's position at the top of the hierarchy of evidence with the ability for confirmatory conclusions, proper evaluation of risks of random errors requires more consideration.

# ASSESSING ASSUMPTIONS FOR STATISTICAL ANALYSES IN RANDOMISED CLINICAL TRIALS

The randomised clinical trial plays a crucial role in evidence-based medicine. Hence incorrect statistical evaluations of trial findings may potentially jeopardise healthcare quality [50]. To guarantee the credibility of trial outcomes and, in some situations to enhance statistical power, most statistical methods require validation of underlying statistical assumptions [50].

## What we showed?

Between January and March 2016, all randomised clinical trials published in six key medical journals (New England Journal of Medicine, The Lancet, British Medical Journal, Journal of the American Medical Association, Annals of Internal Medicine and PLOS Medicine) were identified [50]. These journals were selected based on their high-impact factor and diverse medical specialisations. We incorporated all types of randomised clinical trials regardless of their design, setting, and medical speciality [50]. However, cluster randomised trials were excluded due to the recommended distinctive statistical analysis methods compared to standard randomised clinical trials. Efforts were made to find the protocols and statistical analysis plans for all trials through citations in the main paper, clinical trial registry, or by exploring Google Scholar and PubMed. We extracted data for each trial regarding (1) the nature of the primary outcome, (2) which statistical method was planned for analysing the primary outcome (e.g. in a published protocol or statistical analysis plan), (3) which statistical method was used for analysing the primary outcome, and (4) whether there was a plan (e.g. in a published protocol or statistical analysis plan) to assess the key assumptions of the methods used. In the latter case, we examined (A) whether the results of the assessments of the assumptions were reported and (B) what was done as an alternative strategy if the assumptions were not fulfilled [50].

We identified a total of 83 randomised clinical trials. Four of these 83 trials evaluated two kinds of primary outcomes, leading to 87 primary outcome analyses. Of these analyses, 31 were based on binary data, 21 on continuous data, 31 on time-to-event data, and 4 on count data [50]. In 65 (78%) out of the 83 trials, there was no mention of any evaluation of the underlying assumptions for the applied statistical methods [50].

## Binary data

None (0%) of the 31 trials assessing binary outcomes reported assessments of the underlying assumptions of the statistical methods used [50].

## Continuous data

Only 5 (24%) of the 21 trials assessing continuous outcomes reported assessments of the underlying assumptions. Two of these 21 trials described assessments of the normality of the residual distribution in the main publication, and two other trials reported in the protocol that the normality of the residual distribution would be assessed. Still, the assessment was not reported in the main publication. Using a mixed effects model, one trial had prespecified the assessments of several assumptions, including plots of residuals, deviations from linearity in a regression of continuous data against time, and normality of independence. It stated that alternative analyses would be performed if the underlying assumptions were unmet. This trial, however, did neither specify the nature of the alternative analyses nor report the results of these assessments in the main publication [50].

Several fundamental statistical assumptions exist for the diverse types of statistical methods used to analyse randomised clinical trials. It is rare for trialists to report whether or how these base assumptions were confirmed. While many trialists publish protocols of the trial design and analysis plans that include an outcome hierarchy and a description of the general statistical methodology, it should also be standard practice to report plans for evaluating underlying statistical assumptions [50]. By confirming that no assumption violations are happening, it may be ensured and documented that the correct statistical methodology is applied  [50].

When analysing outcomes of randomised clinical trials, various decisions made during the analysis phase could potentially affect the trial's results; hence, it should also be reported what criteria are used to determine if an assumption has been violated and what actions will be taken if the base assumptions are not met. These steps will minimise the risk of inaccurate or data-driven biased trial conclusions [50]. Without a comprehensive description of the statistical methodology, including evaluations of underlying base assumptions, replicating trial outcomes is difficult or even impossible. We propose that to enhance clarity, thoroughness, and transparency of the reporting of randomised clinical trials, every stage of the trial process, including assessments of underlying base assumptions, must be meticulously described and disclosed [50].

## ASSESSMENT OF ASSUMPTIONS OF STATISTICAL ANALYSIS METHODS IN RANDOMISED CLINICAL TRIALS: THE WHAT, AND HOW

The results of randomised clinical trials and systematic reviews thereof hold, and indeed should hold, the highest place in the hierarchy of evidence [1, 51-54]. Yet, numerous factors may bias these results, including selecting and applying statistical analyses [51, 55-57]. To safeguard the accuracy of research results, and in some instances to enhance statistical power, the verification and validation of underlying theoretical assumptions are often required for most statistical methods [58]. For example, the intervention effect calculated by analysis of covariance (ANCOVA) could be incorrect under certain conditions if residuals are not normally distributed, and the power of ANCOVA often increases when data are log-transformed [59]. Similarly, if normality is ignored and the sample size is small, the Wilcoxon rank sum test may be three to four times more powerful than the independent samples t-test [58].

There is seldom a report by trialists on validating underlying assumptions [60]. Additionally, there is no explicit guideline in the literature on evaluating and reporting these assumptions in the context of a randomised clinical trial [60]. Over the years, there has been a positive trend towards enhancing transparency in medical research, exemplified by initiatives like the CONSORT statement and the EQUATOR network. However, it still needs to be routine to prospectively report the methods employed to evaluate the underlying assumptions of statistical methods used in the analysis of a randomised clinical trial [60]. It is crucial to outline the measures to be taken if the underlying assumptions are not met and the criteria for determining whether an assumption has been violated. Through assumption checks, we aim to ensure that the correct statistical methodology is applied. To promote clarity, thoroughness, and transparency in the reporting of randomised clinical trials, every stage of the trial, including evaluations of assumptions underlying the selected statistical methods, must be meticulously protocolised, detailed, and reported [60]. The objective of this paper was: 1) to assess which underlying assumptions need to be evaluated when using logistic regression, linear regression, and Cox regression in the analysis of results from randomised clinical trials; 2) to assess how to appraise and validate these assumptions; and 3) to consider how to handle violations of these assumptions [61].

Our main emphasis was on randomised clinical trials having two intervention groups, utilising logistic, linear, or Cox regression [60]. These regression analyses allow essential adjustments for stratifying variables used in the randomisation and for baseline covariates [62, 63]. The recommendations in this study were formulated through a two-step process: 1) An exhaustive review of the methodological literature to identify potential assumptions for each method, and 2) a consensus study involving thirteen experts from academic clinical trial centres [61].

“

I. **Systematic survey**: The methodological literature was searched to identify candidate assumptions. Relevant databases (PubMed, Cochrane Library, Google Scholar) were searched using the search terms (assumption, statistical, analysis, randomi*) in February, 2019. Based on the results of the systematic survey, two investigators (TL and JCJ) developed a candidate list of assumptions, and compiled an initial draft of the paper, including:

- General considerations
- Which assumptions to assess
- How to assess if underlying assumptions are violated
- Potential measures in case the assumptions are violated

II. **Consensus study of experts**: The initial draft was distributed to invited selected investigators at different departments and institutions known in our network (see List of co-authors). We applied a Delphi-inspired process focusing on anonymised commenting for the investigators to be un-biased by opinions from other specific investigators. Each investigator at each institution was free to accept, reject, comment, or suggest alternative methods, preferably backed by arguments, results of empirical studies, results of simulation studies, and other references to justify their comments and suggestions. All correspondence went exclusively and one-to-one through an independent facilitator (AKN). AKN collected all comments, assembled them into a report, and wrote a compiled and anonymised summary of the comments. JCJ and TL then commented on the facilitator's report and composed a revised draft of how to test for assumptions. The comments, the revised draft, and the report from the facilitator were then sent to the external investigators for the next round of anonymised comments. This process was repeated six times until all co-authors could accept the final document." [61]

## What we showed

We recommend that a protocol or a detailed statistical analysis plan for a trial should prespecify the key assumptions underlying the chosen statistical analyses, how these assumptions should be assessed, and what should be done if the assumptions are violated [60, 61].

In all regression analyses, assessing for significant interactions between each variable and the intervention factor is advised. The statistician should assess each potential primary interaction between the variables included and the intervention factor. The significance and effect size of each interaction term should be evaluated. The threshold for statistical significance should be determined by the number of tests, with the possibility of using conservative Bonferroni adjusted thresholds (0.05 divided by the number of possible interactions). Additionally, it should be determined whether the interaction is expected to have a significant clinical impact. If the interaction is deemed significant, separate analyses should be presented for each relevant variable (e.g. for each site if there is significant interaction between the intervention factor and 'site'), along with a comprehensive analysis including the interaction term in the model [61].

Various plots, such as histograms or residual plots, can be visually inspected to determine if certain underlying assumptions have been violated. Numerous formal statistical tests are also available to check if certain base assumptions are violated (like the Shapiro-Wilk test, Pearson's X2 test, and the Anderson-Darling test) [64, 65]. Visual inspection has limitations as it requires a subjective evaluation of the plot in question, which may not be consistent or replicable [60, 61]. Formal statistical tests also have their limitations [60, 61]. For instance, tests for normality will often infer that the data is not normally distributed if the data set is large, even if the deviation from normality is insignificant [60, 61]. Conversely, if the data set is minor, serious deviations from normality may not be detected due to limited power and the inherent asymmetry in formal hypothesis testing, as discussed below [60, 61].

We generally advocate for using both graphical plots and formal statistical tests. When differences arise between these two evaluations, the reasons behind these differences, and any subsequent actions, should be meticulously considered and reported [60, 61]. To minimise potential bias, all graphical and formal tests of assumptions should be conducted without the knowledge of the allocated interventions to prevent the data-driven choices of methodology [61]. We recommend that all graphics used to be included in the supplementary materials of the main trial publication. This enables readers to evaluate the adequacy of the methods [61]. Additionally, the process used to determine the choice of methods should be documented [61]. All evaluations and statistical analyses must be conducted, while the statistician is unaware of the randomised treatment allocation. That means intervention groups should be labelled as, for instance, '1' and '2', and any unnecessary variables that could jeopardise the binding of the statistician should be omitted from the dataset used in the main analyses [61]. Ideally, blinded data on all outcomes should be analysed by two statisticians independently. Two independent statistical reports should be submitted to the trial steering committee. If there are inconsistencies between the two reports, potential causes should be pinpointed, and a consensus on the most accurate result should be reached. A final statistical report should be drafted, and all three reports should be published as supplementary material [61].

The viability and validity of presenting a 'cookbook' for statistical analyses can always be challenged. However, evaluating or enhancing a methodology is challenging if it is not thoroughly described. This paper provides comprehensive suggestions on evaluating and addressing potential violations of underlying assumptions for three commonly used statistical methods in the analysis of randomised clinical trial results [61]. Our recommendations are not exhaustive, and adhering to another methodology or systematic plan could also yield valid results, as there are often several valid alternatives when the underlying assumptions are violated. Even so, the current lack of transparency when researchers report how they assess underlying statistical assumptions is noteworthy – in both published protocols and trial publications [60]. Other methodological aspects besides tests for underlying assumptions have been extensively studied for years [56]. Our aim with this paper is to increase awareness of this crucial aspect of methodology, and it should be seen as a supplement to existing recommendations [57, 66].

In evaluating outcomes from randomised clinical trials, the primary aim is usually to determine the effectiveness of a certain intervention. Other trial objectives may require scrutinising assumptions not part of our guidance. For instance, we aim to evaluate the proportional hazard assumption across the compared intervention groups. However, we do not advocate for the evaluation of the proportional hazards assumption for each covariate (for example, by inspecting Schoenfeld residual plots for continuous and categorical covariates [67]) since it is unlikely that violations of this assumption would significantly affect the overall conclusions of the trial regarding the intervention's effectiveness. If hazard ratios for each covariate are essential, then the proportional hazard assumption for that specific covariate should be checked [61].

One potential drawback of our study is that we did not strictly adhere to the Delphi methodology [61, 68]. A Delphi process typically involves multiple live meetings, SKYPE meetings, and telephone conferences for discussing recommendations, which have several benefits, such as in-depth discussions and dedicated project time. However, some researchers may be more influential or charismatic than others or may have a higher degree of fame, which could impede the recommendation of the best methodology if the views or arguments presented by these renowned researchers could be more optimal. Hence, our focus was on anonymous feedback when formulating our recommendations, and we have described our method as a 'Delphi-inspired' approach. Nonetheless, not fully adopting the Delphi process could be seen as a limitation of our study [61]. Our study offers advice on how to validate the assumptions that underlie the commonly used statistical analyses for randomised clinical trials, in addition to providing guidance on identifying and handling violations [61]. We are confident that the credibility of trial outcomes would be enhanced if our recommendations are implemented [61].

# POWER ESTIMATIONS FOR NON-PRIMARY OUTCOMES IN RANDOMISED CLINICAL TRIALS

To circumvent issues with Type I errors (incorrectly rejecting a true null hypothesis) and Type II errors (accepting a false null hypothesis) as well as hasty conclusions from the findings of a randomised clinical trial, it is crucial to 1) limit the number of outcomes [14]; 2) adjust confidence intervals and significance thresholds based on the number of outcome comparisons [14]; and 3) establish an outcome hierarchy (outcomes sorted according to their type and the manner they should be interpreted) [69]. It is advised to predefine the primary and secondary outcomes, detailing how and when they are assessed [70]. To mitigate problems with multiplicity and the interpretation of trial results, it is generally best to assess only one primary outcome and base the sample size on this outcome [14]. The primary outcome in a randomised clinical trial should be the one with the greatest clinical significance to the patients, i.e., a patient-centred outcome. All primary and secondary outcomes in a randomised clinical trial should be either vital for the decision to utilise the intervention or sufficiently validated surrogate outcomes [1, 2, 71]. This article elucidated how to establish a legitimate outcome hierarchy in a randomised clinical trial to minimise issues with Type I and Type II errors, using power estimations of the non-primary outcomes. The central point of the current paper was the overall trial result [69]. Our attention was on binary and continuous outcomes. However, the principles outlined can be applied to most other outcome types [69, 70].

## What we showed

Consider a single randomised clinical trial. If the anticipated sample size has yet to be met, it is crucial to calculate the risks of Type I and Type II errors when making the conclusions of the trial [14, 72]. The threshold for statistical significance (and thus the confidence interval) should be adjusted according to the proportion of the initially planned participants who were randomised [14, 72]. Similar issues can occur with non-primary outcomes when the data are considered inadequate; i.e., when the statistical power is not estimated, it is inappropriate to analyse the data as though it is based on a sufficiently large dataset to assess a minimal important difference [69, 73]. If the estimates for minimal important difference and null effect are both included in the naive 95% confidence interval, it suggests that additional information might be required. Conversely, suppose the minimal important difference estimates are not included in the naive 95% confidence interval. In that case, it is uncertain whether more data are needed to reveal a significant effect or if there is genuinely no significant difference between the groups [14, 69]. When the null effect is not included in the naive 95% confidence interval, and it is uncertain whether there are enough data, it will also be challenging to interpret the analysis results [69]. If there is not enough information, trial results are likely to show misleading results of excessively beneficial or harmful effect estimates [14]. Checking unadjusted naive 95% confidence intervals when the sample size has yet to be met will not be enough as these confidence intervals would be unduly narrow, as previously mentioned [14, 69, 72].

To gauge the statistical power strength of an analysis, it is crucial to determine a minimal significant difference, an incidence in the control group when evaluating a dichotomised result or a standard deviation when assessing a continuous outcome, along with a tolerable risk of Type I error adjusted per number of outcome comparisons [14, 71, 73]. Alternatively, the order of testing secondary outcomes may be predetermined and implemented without further adjustments. Still, it should be halted when the first null hypothesis is not rejected, after which subsequent evaluations become exploratory [14, 69]. Most statistical software can easily calculate both the sample sizes and power estimations of non-primary outcomes [74]. Due to the reasons mentioned earlier, we suggest that during the protocol phase, the statistical power of all non-primary outcomes should be evaluated for confirming or rejecting a minimal important difference [69]. If the power is less than 80% (or 90%), this outcome should be categorised as an 'exploratory outcome' along with the non-validated surrogate outcomes [69, 75]. Alternatively, due to scarce data, the confidence interval and the significance thresholds for the outcome under consideration may need adjustments [14, 69, 72], or the sample size might be re-evaluated and increased so that the power of the non-primary outcome in question reaches 80% (or 90%) [14, 72].

A comprehensive search was conducted for all randomised clinical trials published in the BMJ in 2017, yielding 10 trials. Just one of these trials briefly stated that "A trial of this size will also give more than 80% power to detect important differences in secondary outcome…" [69, 76]. The other nine trials did not discuss the power related to non-primary outcomes, a consideration that is typically overlooked by trialists [69]. As previously discussed, the interpretation of trial results should always consider the intended and obtained sample size. Without power estimations, drawing valid conclusions from non-primary outcome results becomes challenging [69]. Estimating the power of non-primary outcomes is straightforward, yet surprisingly, this is not common amongst trialists. Of course, estimating the power of an outcome comparison necessitates estimating minimal important differences (along with a measure of variance and a tolerable risk of Type I error), which might seem problematic. However, defining minimal important differences for all significant outcomes is essential, even without power estimations, to assess whether statistically significant results have clinical significance for patients [14]. All the required quantities for power estimations (minimal important differences, estimations of proportions in the control group, standard deviations) could potentially be estimated based on a systematic review of studies published before the trial [69].

Reflecting on the clinical significance of outcomes and power estimations is a valuable tool in establishing suitable outcome hierarchies [69]. Beyond determining the necessary sample size, we propose that future trialists consider estimating the power of all non-primary outcomes and even contemplate estimating the power of subgroup comparisons when planning a randomised clinical trial [69]. Power estimations for non-primary outcomes may assist trialists in categorising these outcomes as secondary or exploratory [69]. Power estimations are straightforward, and if utilised systematically, they may help to establish more accurate outcome hierarchies and make trial results easier to interpret [69, 77].

# COUNT DATA ANALYSIS IN RANDOMISED CLINICAL TRIALS

Randomised clinical trials often use two types of count data for evaluating intervention results. The first type includes observations noted as distinct positive figures derived from counting rather than ranking – for instance, the count of severe adverse events or intensive care days [78]. The second type involves tallying events within a certain time frame or a space, counted as event occurrence rates (like the count of adverse events per day) [79]. In some instances (usually where there is a large count), it might be feasible to analyse count data as continuous data (like using ANCOVA). However, it is our experience that count data in randomised clinical trials most often should not be analysed as continuous data because the underlying statistical assumptions are rarely met [80].

Numerous methods exist for analysing count data, but trialists may struggle to select the most suitable one. For instance, the assumptions necessary for choosing the model may not be met, and the model may not adequately fit the data. Additionally, group comparisons via multiple tests yielding varied outcome results can heighten the risk of at least one false positive significant result due to 'play of chance' (type I error), allowing trialists to choose specific tests based on their significance [81, 82]. As a result, a comprehensive procedure for selecting the most trustworthy method for count data analysis should be formulated and published before accessing the trial data [70, 80].

## What we showed

Unlike many observational studies [1], achieving an ideal model fit is not always necessary when analysing count data from randomised clinical trials [80]. The population under investigation in one trial may not match another trial's population, even if they are studying the same disease, intervention, and outcomes. Using data-driven transformations to perfect a model fit in one trial does not guarantee that the same transformations will yield an ideal fit in subsequent trials studying the effects of the same intervention on similar populations. Thus, replicating trial results can be challenging if outcomes or variables are adjusted (e.g. square root, square, or inverse) in various statistical manners to enhance model fit. Moreover, the primary aim of a randomised clinical trial is to determine the effectiveness of an intervention, and precise estimations of variables, coefficients, etc., may not be the main focus. However, to validate trial results and sometimes to maximise their power, the correct choice of statistical method is crucial. The best choice of assumption tests and analysis methods should strike a balance between achieving a model fit and employing a methodology that allows trial results to be replicated and generalised [80]. Like all statistical analyses, the analysis of count data should rely on transparently published, in-depth statistical plans for conducting the analyses [83]. These detailed analysis plans should be developed before data collection – or before researchers or statisticians gain access to the trial data [83]. If these comprehensive statistical analysis plans reveal flaws during the analysis process, then it is crucial that the plans are transparently revised and reported [80, 83].

It is crucial to meticulously consider the design of count data outcomes and their assessments, and it is advisable to publish a detailed statistical analysis plan before the trial results are analysed, ideally before data collection or at least before data access is granted [80]. The chosen model's thorough selection process needs to be outlined in the analysis plan. In the absence of evidence supporting a parametric model, we suggest as standard approaches, adopting either the van Elteren or the $T_{adap2}$ tests and employing bootstrapping to calculate the confidence intervals for the median or mean differences. If more than two trial groups need to be compared, a stratified version of the Kruskal–Wallis one-way analysis of variance test could be applied [80]. The likelihood of biased results when analysing count data in randomised clinical trials may be reduced if our recommendations are followed [80].

## DIRECT-ACTING ANTIVIRALS FOR CHRONIC HEPATITIS C

Around the world, it is estimated that 71 million individuals suffer from chronic hepatitis C, equating to a global prevalence of 1.6% [84, 85]. It is reported by the World Health Organization (WHO) that hepatitis C results in approximately 400,000 deaths per year, primarily due to cirrhosis and hepatocellular carcinoma. In the U.S., the leading cause of chronic liver disease and the most common reason for liver transplants is hepatitis C [86]. Direct-acting antivirals (DAAs), relatively recent medical developments, have been celebrated as a cure for hepatitis C [85, 87]. DAAs function by targeting specific proteins in the hepatitis C virus, which interrupts the virus's replication [85].

Guidelines from leading health organisations, including the American Association for the Study of Liver Diseases (AASLD), European Association for the Study of the Liver (EASL), and WHO, advocate for early DAA treatment for all chronic hepatitis C patients [88-90]. We carried out a systematic Cochrane review, following the eight-step procedure previously outlined [17], in which we searched for all available, published, and unpublished randomised clinical trials that evaluated the effects of DAAs compared with placebo or no treatment for chronic hepatitis C [85]. Our search included databases such as The Cochrane Hepato-Biliary Group Controlled Trials Register, CENTRAL, MEDLINE, Embase, Science Citation Index Expanded, LILACS, BIOSIS, three Chinese databases, Google Scholar, TRIP Database, ClinicalTrials.gov, EMA, WHO International Clinical Trials Registry Platform, FDA, and pharmaceutical company resources. We included trials regardless of their publication type, publication status, and language. We focused on outcomes related to hepatitis C-associated morbidity, serious adverse events, and health-related quality of life as primary outcomes. Secondary outcomes included all-cause mortality, ascites, variceal bleeding, hepato-renal syndrome, hepatic encephalopathy, hepatocellular carcinoma, non-serious adverse events (reported separately), and sustained virologic response. We systematically evaluated bias risks, conducted Trial Sequential Analysis, and adhered to the eight-step procedure to assess statistical and clinical significance thresholds. The overall quality of the evidence was evaluated using the GRADE approach [85].

## What we showed

Our analysis included 138 trials involving a total of 25,232 participants [85]. Most of these trials were short-term and mainly aimed at evaluating the impact of treatment on sustained virologic response. They assessed 51 distinct DAAs, with 128 trials utilising a matching placebo as control intervention. All trials included were deemed to have high risk of bias. Eighty-four trials involving 13,466 participants incorporated DAAs that were either on the market or in the development phase. Fifty-seven trials involved DAAs that were either discontinued or pulled from the market. In 95 trials, the study populations were treatment-naive, 17 trials involved previously treated individuals, and 24 trials consisted of both treatment-naive and pre-treated individuals. The HCV genotypes included were genotype 1 (119 trials), genotype 2 (eight trials), genotype 3 (six trials), genotype 4 (nine trials), and genotype 6 (one trial). We also discovered two ongoing trials [85]. We could not confidently establish the impact of market-available or developing DAAs on our primary outcome of hepatitis C-related morbidity or all-cause mortality. There was a lack of data on hepatitis C-related morbidity and scarce mortality data from 11 trials (DAA 15/2377 (0.63%) versus control 1/617 (0.16%); OR 3.72, 95% CI 0.53 to 26.18, very low-quality evidence). We did not conduct Trial

Sequential Analysis for this outcome [85]. Very low-quality evidence suggested that market-available or developing DAAs did not affect serious adverse events (DAA 5.2% versus control 5.6%; OR 0.93, 95% CI 0.75 to 1.15, 15,817 participants, 43 trials). Trial Sequential Analysis indicated adequate information to reject that DAAs reduce the relative risk of a serious adverse event by 20% when compared with a placebo. Simeprevir was the only DAA leading to a lower risk of serious adverse events when analysed separately (OR 0.62, 95% CI 0.45 to 0.86). However, Trial Sequential Analysis indicated insufficient data to confirm or reject a 20% relative risk reduction. When one trial with an extreme outcome was omitted, the meta-analysis result showed no evidence of a difference [85].

Direct-acting antivirals (DAAs) that were either on the market or in development could potentially reduce the likelihood of not achieving a sustained virologic response, from 54.1% in untreated individuals to 23.8% in those who are treated with DAAs (RR 0.44, 95% CI 0.37 to 0.52, 6886 participants, 32 trials, low-quality evidence). Trial Sequential Analysis supported this outcome. However, only one trial out of 84 evaluated the impact of DAAs on health-related quality of life using the SF-36 mental score and SF-36 physical score [85]. The effect of DAAs that had been discontinued or withdrawn on hepatitis C-related morbidity and all-cause mortality was unclear due to insufficient evidence from trials (OR 0.64, 95% CI 0.23 to 1.79; 5 trials, very low-quality evidence). These DAAs appeared to increase the risk of serious adverse events (OR 1.45, 95% CI 1.22 to 1.73; 29 trials, very low-quality evidence), a conclusion supported by Trial Sequential Analysis [85]. None of the 138 trials offered valuable data regarding the impact of DAAs on other secondary outcomes, such as ascites, variceal bleeding, hepato-renal syndrome, hepatic encephalopathy, or hepatocellular carcinoma [85].

Our study had multiple strengths [85]. We incorporated trials irrespective of publication type and status, the language used, and selected outcomes. We reached out to all pertinent trial authors for more information, if necessary. We applied a systematic review methodology that was predefined and updated, with no modification during the study process [85]. We utilised Trial Sequential Analyses and adjusted our significant thresholds to control random error risks, and we thoroughly evaluated each trial's bias risks assessing systematic error risks. We used the eight-step procedure to determine if the thresholds for statistical and clinical significance were crossed [85]. We also examined the robustness of our results with sensitivity analyses (best-worst, worst-best, etc.) [85]. Last, we documented both aggregate and individual serious adverse events from all included trials that reported them, and any non-serious adverse events were also reported [85].

Our systematic review also had limitations [85].

Our bias risk evaluation revealed that all trials were at high risk of bias. Hence, there is a strong risk that our review results are biased, meaning our findings likely overstate benefits and underplay harms [17, 58, 91-99]. This is the predominant limitation of our review. Trial Sequential Analyses indicated that except for the primary analysis of the impact of DAAs on the risk of serious adverse events, we lacked sufficient data to confirm or reject our anticipated intervention effects. There were not enough trials with an adequate volume of participants evaluating clinically relevant outcomes. It is possible that the numerous neutral meta-analysis outcomes are caused by limited statistical power and that DAAs may indeed have beneficial or detrimental effects.

Additionally, our multiple secondary and subgroup analyses increase the risks of type I errors [17]. Therefore, the risk of random type I errors in this review was substantial. We included all types of DAAs (available or in development) in our primary analysis, and the primary analysis of the impact of DAAs on the risk of severe adverse events indicated that we had enough data to reject a 20% relative risk reduction. It is plausible that different DAAs have varying effects, and including some DAAs in the analysis might have watered down the positive or negative effects of other DAAs. Nevertheless, our analyses detected no signs of heterogeneity, indicating that all the different DAAs appear to have minimal or no clinical impact on the risk of severe adverse events. We primarily centred our attention on the overall combined analysis of DAAs on the market or in development for two reasons: 1. a combined analysis would offer the most considerable statistical power and precision; and 2. it would enable the comparison of different DAAs in subgroup analysis if all types of DAA were included in this current review [85].

A potential limitation was using the composite outcome 'serious adverse events'. By definition, each component of this composite outcome did not necessarily hold the same severity levels, which could have biased the overall outcome result [2]. For instance, if one intervention group experienced more severe adverse events compared to less severe ones in the comparison group, it may have led to overlooking actual severity differences when analysing this composite outcome [2]. The most suitable and patient-relevant primary outcome with minimal methodological limitations would have been all-cause mortality [2]. However, conclusions are rarely drawn from assessing all-cause mortality due to limited sample sizes, which is also evident in our current review. To achieve substantial statistical power, it is often required to use composite outcomes; however, it is crucial always to consider the potential drawbacks of using such outcomes when interpreting review results.

We decided pragmatically to evaluate outcomes at a single assessment time point, specifically, the trials’ results at the most extended follow-up. Most trials only provided short-term results. Therefore, our findings neither confirm nor reject the long-term clinical effects of DAAs, which is another limitation of our current review’s results, especially considering that most detrimental effects of hepatitis C take years to manifest [85].

## Conclusions

The evidence for our main outcomes of interest comes from short-term trials, and we could not determine the effects of long-term treatment with DAAs. The observed hepatitis C morbidity and mortality rates in the included trials were relatively low, and the impact of DAAs on these outcomes remains uncertain. Overall, there is very low-quality evidence suggesting that the current or forthcoming DAAs did not impact serious adverse events. The evidence was inadequate to determine if DAAs positively or negatively impact other chronic Hepatitis C Virus (HCV) clinical outcomes. Simeprevir might have had a positive influence on the risk of serious adverse events. In all other analyses, we could neither confirm nor reject any clinical effects of DAAs.

DAAs might decrease the number of individuals with a detectable virus in their bloodstream. Still, based on current evidence we could not comprehend how sustained virologic response influenced long-term clinical outcomes. Sustained virologic response remains a surrogate outcome that needs proper validation in randomised clinical trials [85]. Our pragmatic approach led us only to evaluate outcomes at one assessment time point, the trials' results at their longest follow-up. Most trials only provided short-term results. Therefore, our findings can neither confirm nor reject any long-term clinical effects of DAAs. This is an additional limitation of our current review results, mainly because most of the detrimental effects of hepatitis C take years to manifest [85].

## DO DIRECT-ACTING ANTIVIRALS CURE CHRONIC HEPATITIS C?

Following the publication of our Cochrane review [85], the British Medical Journal invited us to compose an article addressing ‘uncertainties’ [100]. In this piece, we encapsulated the key findings for medical practitioners and reflected on their clinical implications [100].

Direct-acting antivirals (DAAs), relatively recent additions to the pharmaceutical repertoire, have been lauded as a breakthrough in treating hepatitis C [85, 87]. DAAs inhibit the proteins essential to replicating the hepatitis C virus [85]. The American Association for the Study of Liver Diseases (AASLD), European Association for the Study of the Liver (EASL), and the World Health Organization (WHO) all advocate for the early administration of DAAs to all chronic hepatitis C patients [88-90]. According to these guidelines, treatment is deemed successful if it yields a sustained virological response, defined as the absence of detectable hepatitis C virus RNA in the blood 12-24 weeks post-treatment and beyond [88-90]. Nevertheless, the clinical consequences of achieving a sustained virological response remain uncertain [85, 100]. The rationale for using sustained virological response as an indirect indicator of reduced mortality, lowered risk of liver cancer, and fewer liver-related complications are solely based on observational studies. These studies are often not controlled and always susceptible to confounding factors [101-103]. Describing it as a “cure” is misleading as some patients who have exhibited sustained virological response may experience a relapse with genetically identical viruses, indicating that the virus may have been dormant in their bodies. Furthermore, those achieving a sustained virological response could still develop end-stage liver disease [100, 104]. The question remains whether DAAs provide tangible benefits for chronic hepatitis C patients in terms of lessening the clinical risks associated with hepatitis-related complications and mortality [100].

## What we showed

We drew data from our Cochrane review [85] involving an exhaustive search of all current, published, and unpublished randomised clinical trials examining the effects of DAAs versus placebo or no intervention for chronic hepatitis C. We searched The Cochrane Hepato-Biliary Group Controlled Trials Register, CENTRAL, MEDLINE, Embase, Science Citation Index Expanded, LILACS, and BIOSIS; three Chinese databases, Google Scholar, TRIP Database, ClinicalTrials.gov, EMA, WHO International Clinical Trials Registry Platform, FDA, and pharmaceutical company sources. The Cochrane systematic review of 138 randomised clinical trials (with 25,232 participants) conducted in 2017 assessed 51 distinct DAAs against placebo or no intervention [85]. We considered adults suffering from chronic hepatitis C, irrespective of gender, ethnicity, job, place of living, infection tenure, and disease progression. Both those new to treatment and those with prior treatment experience were included [85]. Of all included trials, eighty-four involved DAAs that were either available on the market or were still in the development phase (13,466 participants) [85]. Fifty-seven trials were on DAAs that have since been discontinued or withdrawn from the market [85]. The follow-up spanned from 1 week to 120 weeks, averaging 34 weeks. All trials and outcome results were at a high risk of bias [85]. Most trials primarily evaluated the effect on sustained virological response with limited data on clinically significant outcomes and none on long-term effects [85]. A meta-analysis of the effects of all DAAs available or in the developmental phase showed no evidence of a difference in all-cause mortality in DAA recipients compared to controls (2996 participants, 11 trials, very low-quality evidence) [85]. The number of patients with hepatitis C morbidity and mortality observed in the trials was low, and the effects of DAAs on these outcomes were unclear [85]. Meta-analysis and Trial Sequential Analysis indicated that DAAs compared with placebo or no intervention did not appear to affect the risk of serious adverse events (any serious clinical adverse event defined according to ICH-GCP (e.g. death, hospitalisation, persisting adverse events) [85, 105]. More patients achieved sustained virological response with DAAs compared to controls (6886 participants, 32 trials, low-quality evidence) [85]. However, there was no evidence to determine the effects of DAAs on clinically important outcomes such as ascites, variceal bleeding, hepato-renal syndrome, hepatic encephalopathy, and hepatocellular carcinoma. No blinded trials on health-related quality of life were found [85].

## Clinical implications

We encourage physicians to communicate openly with patients about the indefinite long-term health outcomes, potential adverse effects, and financial implications associated with DAA treatment. Patients should be informed that while these medications are likely to eradicate the virus from their bloodstream, there is no proven evidence that DAAs lower the threat of long-term liver complications. The future possibility of cirrhosis, cancer, or the need for a liver transplant cannot be ruled out despite treatment with DAAs [100]. Patients need to be aware of the steps that can be taken to reduce the likelihood of spreading the virus, such as avoiding risky injection methods and unsafe blood transfusions [100]. Additionally, patients need to reduce activities linked with the fast progression of liver disease, like alcohol consumption, drug misuse, and weight gain [100, 106]. The main reason for liver transplantation is end-stage cirrhosis caused by hepatitis C. After a liver transplant, reinfection is inevitable, and at least 70% of patients develop chronic liver disease within three years [107]; the effectiveness of DAAs in these circumstances remains unclear [100].

## SELECTIVE SEROTONIN REUPTAKE INHIBITORS VERSUS PLACEBO IN PATIENTS WITH MAJOR DEPRESSIVE DISORDER. A SYSTEMATIC REVIEW WITH META-ANALYSIS AND TRIAL SEQUENTIAL ANALYSIS

Selective serotonin reuptake inhibitors (SSRIs) are commonly the first-line treatment for depression, and their prescription rates have been on the rise [108, 109]. Several reviews with meta-analysis have evaluated the impact of SSRIs on adults suffering from major depressive disorder [110-115], concluding that SSRIs help alleviate depressive symptoms [110-115]. However, these reviews have been limited by not using predefined Cochrane methodology [110-115], only focusing on certain subgroups of patients with depression [92, 93], not thoroughly searching all pertinent databases [110-116], not systematically evaluating the potential harms [86-91, 93], and not systematically assessing risks of bias [110-116]. Therefore, previous evidence regarding the effects of SSRIs remained uncertain [117].

Following the eight-step procedure as outlined earlier [17], we carried out a systematic review where we searched for all current, published, and unpublished randomised clinical trials that evaluated the effects of SSRIs against placebo, 'active' placebo, or no intervention among adults diagnosed with major depressive disorder [117]. We searched for eligible randomised clinical trials in various resources like The Cochrane Library's CENTRAL, PubMed, EMBASE, PsycLIT, PsycINFO, Science Citation Index Expanded, clinical trial registers of Europe and the USA, pharmaceutical companies' websites, the U.S. Food and Drug Administration (FDA), and the European Medicines Agency up until January 2016. A minimum of two independent investigators extracted all the data. We also utilised Cochrane systematic review methodology, Trial Sequential Analysis, and the calculation of the Bayes factor. The primary outcomes considered were depressive symptoms, remission, and adverse events. Secondary outcomes included suicides, suicide attempts, suicide ideation, and quality of life. We systematically assessed risks of bias, performed Trial Sequential Analysis, and adhered to the eight-step procedure to assess statistical and clinical significance thresholds. The overall quality of the evidence was assessed using GRADE [117].

## What we showed

We included 131 randomised placebo-controlled trials involving 27,422 participants in total [117]. Notably, no trials used 'active' placebo or no intervention as controls. All the trials were at a high bias risk.

SSRIs significantly reduced the Hamilton Depression Rating Scale (HDRS) at the end of treatment (mean difference −1.94 HDRS points; 95% CI −2.50 to −1.37; $P<0.00001$; 49 trials; Trial Sequential Analysis-adjusted CI −2.70 to −1.18); Bayes factor below a predefined threshold ($2.01*10^{-23}$). The effect estimate, however, was below our predefined threshold for clinical significance of 3 HDRS points.

SSRIs significantly decreased the risk of no remission (RR 0.88; 95% CI 0.84 to 0.91; $P<0.00001$; 34 trials; Trial Sequential Analysis adjusted CI 0.83 to 0.92); Bayes factor (1426.81) did not confirm the effect).

SSRIs significantly increased the risks of serious adverse events (OR 1.37; 95% CI 1.08 to 1.75; $P=0.009$; 44 trials; Trial

Sequential Analysis-adjusted CI 1.03 to 1.89). This translates to 31/1000 SSRI participants experiencing a serious adverse event compared to 22/1000 control participants. Furthermore, SSRIs notably increased the risk of several non-serious adverse events.

There was a significant lack of data on suicidal behaviour, quality of life, and long-term effects.

Our current systematic review had several strengths. Our protocol was registered before the systematic literature search was performed, and we searched all relevant databases. Independent authors performed double data extraction, minimising the risk of incorrect data extraction. We used Trial Sequential Analysis to control the risks of random errors. The primary outcome analyses demonstrated that the collected information sizes were adequate. Limited signs of statistical heterogeneity were shown in both visual assessments of forest plots and statistical tests. Therefore, our review results are enhanced in validity, suggesting that the effects shown are consistent across different trials. Several prior reviews and meta-analyses have evaluated the effects of SSRIs, concluding that SSRIs significantly impact depressive symptoms. Our current results align with the estimated results of these reviews and meta-analyses, suggesting that SSRIs only benefit patients with a few HDRS points. This strengthens the validity of our current results. Additionally, we detailed the risks of serious and non-serious adverse events, discovering that both were significantly increased by SSRIs [117].

Our systematic review also had several limitations [117]. The mean differences in our HDRS were averaged effects. Consequently, it is not accurate to assert that SSRIs do not exert medically important effects on every participant with depression.

For instance, certain patients with severe depression compared to those with mild depression (like so-called professional patients or symptomatic volunteers) could potentially gain from SSRIs, even though there is no empirical evidence to support this theory. However, this ‘constraint’ is inherent in any clinical research finding. Specific patients may benefit from a particular intervention, even when credible research findings have demonstrated that this intervention ‘generally’ is ineffective or harmful. All the trials were at a high risk of bias across multiple bias risk domains, and especially the risk of incomplete data, selective reporting of outcomes, and insufficient blinding bias could bias our review findings [117]. Our GRADE evaluations showed that the quality of the evidence should be considered very low primarily due to the high bias risks. These high risks of bias cast doubt on the reliability of our meta-analysis outcomes, as trials with a high bias risk tend to overestimate benefits and underestimate harms [117]. SSRIs’ ‘true’ effect may not even be statistically significant [117].

SSRIs are thought to alter the levels of critical neurotransmitters in the brain and thereby could influence depressive symptoms. Yet, whether these impacts are advantageous and hold clinical significance is uncertain. Determining a meaningful threshold for clinical importance is challenging, and evaluating clinical significance should ideally go beyond a threshold on an outcome scale [118]. Major depressive disorder interferes with daily life, increases the likelihood of suicidal behaviour, and reduces the quality of life [119]. So, certain adverse effects might be tolerable if SSRIs have clinically significant beneficial results [1, 52, 119]. Hence, we predefined a threshold for clinical significance and evaluated the trade-off between beneficial and harmful effects [1, 52, 117, 120].

The clinical significance threshold we chose was a difference of 3 points between drug and placebo on the 17-item HDRS scale (which ranges from 0 to 52 points) or a standardised mean difference effect size of 0.50 [117]. This standard is recommended by the National Institute for Clinical Excellence (NICE) in England and has been used in other reviews [111, 121, 122]. However, there is no universal acceptance of these recommendations, and they have been critiqued [110]. Some suggest the following interpretations of standardised mean difference: 0.2 is a small effect, 0.5 is a moderate effect, and 0.8 is a large effect [123, 124]. A study shows that a difference of up to three points between SSRI and placebo on the HDRS scale represents 'no clinical change' [125]. Another credible study indicated that a 3-point difference between SSRI and placebo is not noticeable to clinicians, and a difference of 7 HDRS points, or a standardised mean effect size of 0.875, is needed to represent 'minimal improvement' [126].

It has been theorised that the 'placebo' response in antidepressant trials has been increasing in recent years [127]. If a 'response' to a placebo exists, it should be considered when interpreting the mean difference between a drug and a placebo. However, it is doubtful that depressed patients experience a significant placebo effect [128], and recent evidence shows that the placebo response has remained constant for the past 25 years [127]. Even considering our rather small predefined minimal thresholds for clinical significance, the impact of SSRIs on depressive symptoms was not clinically significant. Moreover, our meta-analyses found that SSRIs significantly increase the risk of both serious and non-serious adverse events [117]. In 2009, the Committee for Medicinal Products for Human Use (CHMP) concluded "……… that, as no public health concerns have been identified, no regulatory action is necessary on the basis of Kirsch et al.'s findings" when the team questioned the benefits of antidepressants [118]. According to our research, we now think there is credible proof for public concern regarding the effects of SSRIs. We concur with Andrews et al. that antidepressants appear to cause more harm than good [129]. We have demonstrated that SSRIs significantly heighten the risks of serious and non-serious adverse events. The observed harmful effects appear to surpass the potential minor beneficial clinical effects of SSRIs, if any exist. Our findings corroborate the results from other studies questioning the efficacy of SSRIs [122, 130], yet contrast with the conclusions of other reviews that posit SSRIs as effective treatments for depression [110, 113, 116, 131]. However, our current analyses offer the most thorough systematic review on the subject, and we hope it will guide clinical practice [117].

## Conclusions

While it appears that SSRIs have a statistically significant impact on symptoms of depression when compared to placebo, the clinical relevance of these effects remains dubious, and all trials were at c high risk of bias. Additionally, using SSRIs compared to placebo notably heightens the risk of serious and non-serious adverse events. Our findings suggest that the detrimental effects of SSRIs for major depressive disorder, compared to placebo, seem to surpass any possible minor benefits [117].

## SHOULD ANTIDEPRESSANTS BE USED FOR MAJOR DEPRESSIVE DISORDER?

The World Health Organization (WHO) approximates that over 300 million people worldwide suffer from major depressive disorder, marking it the principal cause of global disability [132]. The lifetime occurrence of this disorder lies between 10% and 20% [133, 134]. Antidepressants are frequently employed to manage depression. The utilisation of antidepressants is extensive, particularly in the Western world, and their use is escalating in many countries [135]. A 2017 report from the National Health and Nutrition Examination Survey revealed that from 2011 to 2014, approximately one in eight individuals aged 12 and older in the U.S. reported using antidepressants in the preceding month [136]. Over 15 years, antidepressant usage surged by almost 65% [136], and over 60% of U.S. residents on antidepressants have been using them for over two years [136]. Guidelines from the United Kingdom National Institute for Health and Care Excellence (NICE), the American Psychiatric Association, and other authorities [137-143] recommend antidepressants for treating major depressive disorder, either as monotherapy or combined with psychotherapy. Internationally, both psychiatrists and general practitioners routinely prescribe antidepressants for depression. Various types of antidepressants are available [144], with selective serotonin reuptake inhibitors (SSRIs) being the most popularly prescribed and often chosen as first-line treatment for depression [145]. We conducted a narrative review of the evidence regarding the effectiveness of antidepressants compared to placebo for patients diagnosed with major depressive disorder [146]. Two researchers independently explored the Cochrane Library, BMJ Best Practice, and PubMed till June 2019 using the search terms “depression” and “antidepressants”, focusing on narrative and systematic reviews published in English since 1990. We included any review that evaluated the benefits and harms of any antidepressant compared to placebo in adults. We also scrutinised references of the identified articles and incorporated relevant guidelines' recommendations where applicable.

## What we showed

In clinical research, the impact of antidepressants on depression is typically evaluated by measuring their influence on the severity of depressive symptoms using a scale like the Hamilton depression rating scale (HDRS) with 17 items, ranging from 0 to 52 points. The HDRS is widely acknowledged by psychiatrists globally as the primary depression rating scale. The National Institute for Health and Care Excellence (NICE) previously proposed that a difference of three points on the HDRS or a standardised mean difference (SMD) of 0.5 met the criteria for clinical significance or minimal important difference [147]. However, NICE no longer features these thresholds for clinical significance on its website, and these thresholds have faced criticism. Despite this, several studies assessing antidepressants have used these thresholds [111, 122]. In addition, the 0.5 SMD threshold, initially introduced by Cohen to indicate a 'moderate' effect, has been employed as a minimum important difference in numerous studies across various medical fields [148].

It is critical to note that the thresholds proposed by NICE were not based on empirical evidence and are presumably underestimated. Research indicates a mean difference of up to three points on the HDRS between SSRI and placebo corresponds to 'no change' in the patient's condition. Consequently, a more stringent criterion for clinical significance has been advocated. It is approximated that a 'minimal improvement' corresponds to a seven-point change on the HDRS or an SMD of 0.875 [125]. These have been suggested as empirically derived thresholds for a minimal important difference [149]. However, these empirically derived thresholds do not necessarily represent what patients perceive as the slightest beneficial effect of antidepressants. Nevertheless, these proposed thresholds for clinical significance should be considered when analysing depression research results. As stated, all existing evidence indicates that a minor change, for instance, two HDRS points, should be regarded as a minimal effect and is likely imperceptible to the average depressed patient [125]. Although the available evidence suggests that the average effect on depressive symptoms is minimal, theoretically and in clinical practice, some patients might significantly benefit from antidepressants. Yet, suppose the average effect is minimal and near zero, and some patients significantly benefit. In that case, a similar number of patients must be significantly harmed by antidepressants to maintain the average effect close to zero. Furthermore, no studies have definitively determined which patients will respond positively to antidepressants and which will not [146].

Researchers frequently convert the HDRS scale into a binary score, for example, distinguishing between responders and non-responders based on a criterion of 50% or more improvement on the HDRS from baseline. However, this conversion of continuous data into two categories, known as dichotomisation, has been proven to be problematic due to various methodological constraints, and it often leads to biased outcomes [150, 151]. Interestingly, someone who improves by 50% or more is classified as a responder, and someone who improves by just 49% is considered a non-responder, thus exaggerating the perceived difference between these individuals [151]. Similarly, a person with an improvement of 50% is seen as identical to someone whose symptoms have entirely vanished, and  someone with a 49% improvement is treated the same as someone without improvement. How the data are distributed will have an impact. Still, even if more participants in the antidepressant group cross the arbitrary cut-off point compared to the control group, the HDRS difference might be minimal. Further, if the data distribution or cut-off points differ, a real difference between groups might not be detected when evaluating these dichotomised outcomes. Therefore, when analysing these dichotomised outcomes, there is a significant risk of both overestimating the benefits and missing a 'true' effect. As a result, binary outcome results such as 'response' or 'remission' should not be used to determine statistical or clinical significance and should be interpreted with caution [146].

The validity of the HDRS as an 'interval scale', where the distance between any two successive points should be identical irrespective of where you are on the scale, has been challenged [152]. It is important to mention that when different evaluation scales are employed, such as the Beck Depression Inventory, Montgomery–Åsberg Depression Rating Scale, or Hamilton 6-item scale, these findings align with those of the HDRS, leading to statistically significant results that may not be relevant to the average patient [117]. The HDRS might be better classified as an ordinal scale, where the distance between any two points cannot be considered equal to the distance between two different points elsewhere on the scale [152].

The HDRS has been criticised for being both conceptually and psychometrically deficient [153]. As a result, there could be a case that it is impracticable to determine the clinical significance of a specific HDRS score and that the foundational evidence is essentially flawed due to the reliance on the HDRS. To prove that antidepressants provide more benefits than harms, new studies using more clinically relevant outcome scales and superior designs are necessary before these drugs can be recommended for depression treatment [146].

## How good are antidepressants?

Numerous reviews have evaluated the impact of antidepressants versus placebos in treating depression [91]. These reviews consistently demonstrate that antidepressants have a statistically significant effect on depressive symptoms. Most of these reviews, however, were not systematic reviews as per PRISMA guidelines but narrative reviews [154]. The following paragraphs will outline two of the most extensive and recent systematic reviews [154].

In 2017, our group conducted a systematic review using the previously established eight-step procedure [17]. This process involved searching all relevant databases, systematically assessing both the beneficial and harmful effects, and performing a predefined evaluation of the clinical significance of antidepressants [117]. Like all preceding ones, the review found that SSRIs had a statistically significant impact on depressive symptoms when compared to placebo [117]. However, the effect size of SSRIs (1.94 HDRS points; 95% CI -2.50 to -1.37 or -0.23 SMD; 95% CI -0.31 to -0.14) fell short of the pre-defined threshold for clinical significance (as per the NICE criteria mentioned earlier) and was far from ‘minimal improvement’ (e.g. a seven-point difference on the HDRS or an SMD of 0.875). Trials examining long-term effects suggested that these effects were even less pronounced than short-term effects. The trials provided virtually no data on suicidal tendencies or quality of life. We discovered that SSRIs significantly heightened the risk of both serious and non-serious adverse events. All the trials included in the review were at a high risk of bias, suggesting a high likelihood that the review's results overestimated the beneficial effects and underestimated the harmful effects of SSRIs [58, 97, 99]. We concluded that the potential minor benefits of SSRIs appeared overshadowed by their harmful effects.

Our GRADE evaluation of the evidence quality: very low quality [117].

The Lancet recently published a comprehensive network meta-analysis [155]. The review included placebo-controlled and head-to-head trials of 21 popular antidepressants, including SSRIs. The study assessed all outcomes as closely to 8 weeks as possible, meaning only short-term results were considered. The authors also evaluated 'acceptability' (measured by the number of patients who discontinued treatment for any reason) and the fraction of patients who quit early due to adverse effects. However, these results are challenging to interpret clinically – as patients may persist with antidepressants despite experiencing adverse effects, and there might be other reasons than adverse effects as a cause for quitting. The network meta-analysis did not assess serious or non-serious adverse events. The findings relating to benefits were practically identical to previous reviews, indicating that antidepressants seem to lower depressive symptoms with a statistically significant effect (SMD 0.30; 95% credibility interval 0.26 to 0.34), but such an effect size is likely clinically irrelevant [149]. The baseline depression severity was high (HDRS 25.7). One major limitation of the review was that only 18% of the included trials were at low risk of bias, which implies a high risk that the review's results overstate the beneficial effects and downplay the harmful effects of antidepressants [58, 97, 99].

Our GRADE evaluation of the evidence's quality: very low quality.

As noted, numerous other reviews have been published, mostly affirming that antidepressants have statistically significant effects. Still, only a handful of these reviews evaluated the clinical significance of the review results. Yet, if the effect estimates from the previous reviews are compared to, for example, the NICE criteria, then the previous reviews verify that antidepressants generally have minimal positive impact on depressive symptoms. Despite the lack of solid evidence supporting the beneficial effects of antidepressants for depression, we cannot rule out that there is evidence supporting beneficial effects for other conditions besides major depressive disorder.

The primary limitation of the existing evidence on the impact of antidepressants is that most past studies that evaluated these effects have a high or unclear risk of bias [97, 99, 117, 155]. Even when trials use a matched placebo, patients may deduce whether they are receiving an antidepressant or a placebo due to identifiable adverse effects in the intervention group and their absence in the control groups, potentially undermining the blinding process and the accurate evaluation of subjective symptoms. Studies with high or unclear risk of bias often overestimate benefits and underestimate potential harms [58, 91-99]. Despite this bias leading to an overestimation of the beneficial effects in review results, the difference between antidepressants and placebos on depressive symptoms is minimal, and the 'true' effect of antidepressants might not be statistically significant.

The evidence for antidepressants also suffers from limited applicability due to the selection of specific patient groups. The results from a large-scale trial carried out in a clinical environment (the StAR*D trial) revealed that conventional clinical trials would exclude around 77.8% of the participants in the StAR*D trial [156]. The patients in this trial showed a marginal improvement in the HDRS after three months of treatment with an SSRI, compared to a significant improvement in conventional comparator trials, which, similar to the StAR*D trial, do not include placebo controls [157]. This suggests that in a clinical context, the benefits of antidepressants are minimal and that the exclusion criteria commonly used in randomised clinical trials result in exaggerated effect estimates [146].

Lundh and colleagues demonstrated that when the manufacturing company is involved, it leads to more positive results and conclusions than when sponsored by other sources [95]. Studies funded by the industry more frequently reported benefits, with a relative risk (RR) of 1.27 and more often yielded positive conclusions with an RR of 1.34 [95]. Ebrahim and team identified 185 suitable meta-analyses that evaluated the effects of various antidepressants. In those meta-analyses that included an author who was an employee of the drug's manufacturer, there was 22 times less likelihood of negative comments about the drug compared to other meta-analyses [91]. The systematic review mentioned previously, which evaluated the effects of SSRIs, deduced that a significant majority of the included trials (39 out of 43 studies) that provided useful information were at a high risk of being influenced by 'for profit' bias [117]. The review did not find any significant statistical effect of SSRIs on HDRS in studies at low risk of 'for profit' bias. However, positive effects were statistically significant in trials with a high or unclear risk of 'for profit' bias [94] [117]. Considering that a significant amount of research on antidepressants is susceptible to 'for profit bias', it is likely that past results may have overestimated benefits and underestimated harms [91, 95]. This should be considered when interpreting available research findings [146].

A small individual patient data meta-analysis found evidence for a correlation between the severity of depression and the benefit derived from the treatment [111], but three larger individual patient data meta-analyses did not find any such correlation [110, 158, 159]. Some other studies did find a correlation between the severity of depression and treatment benefit [122], but even for the most severely depressed patients, the effects were minimal [122]. Consequently, there is no definitive evidence to suggest that antidepressants would be more beneficial for patients with severe depression than those with mild or moderate depression [146].

Most studies and reviews have only evaluated the short-term effects of antidepressants, typically ranging from four to eight weeks [146]. The long-term effects of these drugs remain uncertain. Data are scarce on the long-term effects of antidepressants, such as after a year. A recently conducted review assessing outcomes after 24 weeks revealed that the long-term effects of antidepressants (SMD 0.34) are as modest as the short-term effects [160]. A clinical practice guideline from NICE showed similar findings (SMD 0.28) [147]. It is plausible that long-term antidepressant treatment could potentially deteriorate outcomes [161]. In light of the lack of evidence for benefits, no substantiated evidence supports long-term treatment with antidepressants [146].

## What are the harms?

Research has indicated that SSRIs, the most frequently prescribed antidepressant, increase the risks of serious and non-serious adverse events [117]. Although the relative risk of serious adverse events is fairly high, the absolute risk remains small, while non-serious adverse effects of SSRIs are more prevalent. The most concerning adverse effects of prolonged SSRI use tend to be digestive issues, sleep irregularities, and sexual dysfunction [162], with the latter potentially persisting even after discontinuing treatment [163]. Also, there appears to be an elevated risk of birth abnormalities in babies born to women who were administered certain SSRIs during their pregnancy [146, 164]. The adverse effects of other antidepressants, such as serotonin-norepinephrine reuptake inhibitors (SNRIs) and tricyclic antidepressants (TCAs), have not been extensively studied in systematic reviews. Still, they could potentially be more severe [146]. Non-randomised studies have demonstrated, for example, that TCAs can lead to seizures and even death due to slowed intraventricular conduction, resulting in complete heart block or ventricular arrhythmias [162]. Furthermore, as the existing evidence is based on short-term trials, it is plausible that the current estimates of the adverse effects of antidepressants may be underreported [162]. Relying on short-term results is generally problematic, as many patients receive long-term antidepressant treatment [136, 146].

Post-SSRI withdrawal symptoms typically appear within days of stopping the medication and can last several weeks, even with a gradual reduction. However, these symptoms may include late onset and lasting disturbances and could be misinterpreted as early signs of relapse [165]. There are notable similarities between the withdrawal symptoms of SSRIs and other antidepressants like venlafaxine and duloxetine [165]. A recent review discovered that a significant number of individuals who exhibit withdrawal symptoms post-antidepressant treatment experience them for over two weeks, and it is not rare for withdrawal to last several months [166]. Withdrawal symptoms can potentially be eased by resuming the antidepressant that caused the symptoms, making it difficult for some people to stop taking the medication once started [165]. This could also explain why some research suggests a decreased relapse risk when continuing antidepressants versus discontinuing them [167, 168]. The withdrawal symptoms might be why patients who discontinue antidepressants might do worse than those who remain on them [146].

## Combination of antidepressants and psychotherapy

Antidepressants in combination with psychotherapy are recommended for major depressive disorder by NICE, the American Psychiatric Association, and various other guidelines [140, 141, 143, 169-171]. A non-systematic review indicated that combining antidepressants with psychotherapy had a statistically significant impact. However, the effect was minor and did not meet the NICE standards (SMD 0.35; 95% CI 0.24 to 0.45; P0.001) [137]. Over an extended follow-up period, there was no noticeable difference between psychological treatments and those combined with medication [137]. Two additional reviews [172, 173] supported this finding. The benefits of incorporating antidepressants into psychotherapy appear minimal, similar to the impact of antidepressants when used as a standalone treatment [146].

## Conclusions

Current evidence suggests that the disadvantages of antidepressants outweigh the benefits [146]. We need comprehensive randomised clinical trials with low bias risk, which incorporate an ‘active placebo’ (a placebo that mimics the adverse effects of the treatment, leading the patient to believe they are receiving an actual treatment). These trials should not only assess depression symptoms and quality of life but also systematically evaluate potential adverse effects, with long-term follow-up included. They should be designed to conclusively establish whether antidepressants increase risks such as suicides, hospitalisation, mortality etc. (the precise sample size would be determined by the occurrence rate of such events in the control group).

## DRUG-ELUTING STENTS VERSUS BARE-METAL STENTS FOR ACUTE CORONARY SYNDROME

Cardiovascular disease, with ischaemic heart disease as the major subset, is globally recognised as the leading cause of death [172]. As per the World Health Organization (WHO), 7.4 million fatalities were attributable to ischaemic heart disease in 2012, accounting for 15% of worldwide deaths [173]. Ischaemic heart disease's prevalence and treatment costs are rising due to increased longevity and reduced mortality rates [174]. We carried out a systematic Cochrane review following the previously outlined eight-step procedure [17], in which we sought all ongoing, published, and unpublished randomised clinical trials that evaluated the impact of drug-eluting stents versus bare-metal stents in patients with acute coronary syndrome [175]. We explored randomised clinical trials in databases such as the Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, Embase, LILACS, SCI-EXPANDED, and BIOSIS, covering the period from their inception until January 2017. Additionally, we investigated two clinical trial registers, the European Medicines Agency and the US Food and Drug Administration databases, and pharmaceutical company websites. We also reviewed the reference lists of review articles and relevant trials. We considered trials for inclusion regardless of their publication type, status, date, or language. Our primary outcomes were all-cause mortality, major cardiovascular events, serious adverse events, and quality of life. Our secondary outcomes were angina, cardiovascular mortality, and myocardial infarction. Our primary assessment time point was at maximum follow-up. We systematically evaluated the risks of bias, performed Trial Sequential Analysis, and adhered to an eight-step process to determine statistical and clinical significance thresholds. We assessed the overall quality of the evidence using GRADE [175].

## What we showed

Our study incorporated 25 trials randomising 12,503 participants in total [175]. All trials were at high risk of bias, and the evidence quality, according to GRADE, ranged from low to very low [175]. We included 22 trials in which the participants had ST-elevation myocardial infarction, one trial where the participants had non-ST-elevation myocardial infarction, and two trials where the participants had various acute coronary syndromes [175].

Upon analysing all-cause mortality and major cardiovascular events at maximum follow-up, there was no evidence of a difference when comparing drug-eluting stents to bare-metal stents. The absolute death risk was 6.97% in the drug-eluting stents group versus 7.74% in the bare-metal stents group, as per the risk ratio (RR) of 0.90 (95% confidence interval (CI) 0.78 to 1.03, 11,250 participants, 21 trials/ 22 comparisons, low-quality evidence). The absolute risk of a significant cardiovascular event was 6.36% in the drug-eluting stents group versus 6.63% in the bare-metal stents group, as per the RR of 0.96 (95% CI 0.83 to 1.11, 10,939 participants, 19 trials/ 20 comparisons, very low-quality evidence). However, our Trial Sequential Analysis indicated insufficient data to confirm or reject the anticipated 10% risk ratio reduction in all-cause mortality or major cardiovascular events at maximum follow-up [175].

Upon analysing serious adverse events at maximum follow-up, there was evidence of a benefit when comparing drug-eluting stents to bare-metal stents. The absolute risk of a serious adverse event was 18.04% in the drug-eluting stents group versus 23.01% in the bare-metal stents group, as per the RR of 0.80 (95% CI 0.74 to 0.86, 11,724 subjects, 22 trials/ 23 comparisons, low-quality evidence), and our Trial Sequential Analysis confirmed this result. When we looked at each specific type of adverse event included in the serious adverse event outcome separately, most were target vessel revascularisation. When target vessel revascularisation was analysed independently, meta-analysis showed evidence of a benefit of drug-eluting stents, and our Trial Sequential Analysis confirmed this result [175].

Upon analysing cardiovascular mortality and myocardial infarction at maximum follow-up, there was no evidence of a difference when comparing drug-eluting stents to bare-metal stents (RR 0.91, 95% CI 0.76 to 1.09, 9248 participants, 14 trials/ 15 comparisons, very low-quality evidence/ RR 0.98, 95% CI 0.82 to 1.18, 10,217 participants, 18 trials/ 19 comparisons, very low-quality evidence). Trial Sequential Analysis indicated that we lacked sufficient data to either confirm or reject our anticipated risk ratio reduction of 10% on cardiovascular mortality and myocardial infarction [175]. No trials assessed quality of life or angina [175].

There are numerous strengths to our review [175]. We incorporated trials without considering their language of publication or the outcomes assessed. We reached out to all pertinent authors for additional information if needed. Our review included more participants than any prior systematic review, enhancing our ability to detect significant differences between the intervention and control groups. We adhered to our pre-published, peer-reviewed protocol and conducted our review using Cochrane's recommended methods and additional methodological research findings [175]. We carried out Trial Sequential Analyses and applied the eight-step procedure to determine whether the statistical and clinical significance thresholds were crossed [17]. This enhanced the solidity of our outcomes and conclusions. We also conducted sensitivity analyses (best-worst case and worst-best case) to verify the validity of our results. Our meta-analyses exhibited minor statistical heterogeneity, which bolsters the credibility of our results [175].

Our systematic review also has some limitations [175]. The quality and quantity of the included trials impact our findings, interpretations, and conclusions. The relatively brief follow-up period reported could obscure potential differences; a more extended follow-up might uncover such differences. A common oversight, and another limitation, is that we have compared drug-eluting stents with bare-metal stents without definitive evidence that the latter is statistically and clinically more beneficial than mere balloon dilation or no intervention. Although indications suggest this, conclusive evidence from systematic reviews, meta-analysis, and Trial Sequential Analysis, considering bias risks, is still lacking [2].

## Conclusions

The current evidence suggests that drug-eluting stents could potentially result in fewer serious adverse events compared to bare-metal [175]. Trial Sequential Analysis revealed that there currently is not sufficient data to assess a 10% decrease in risk ratio for all-cause mortality, major cardiovascular events, cardiovascular mortality, or myocardial infarction. Also, information needs to be provided concerning quality of life or angina. The evidence assessed in this review ranged from low to very low in quality, suggesting that the 'true' effects might significantly deviate from the results discussed. There is a need for more randomised clinical trials with low risks of bias and low risks of random errors to correctly evaluate the benefits and harms of drug-eluting stents for acute coronary syndrome.

Additional data on all-cause mortality, major cardiovascular incidents, quality of life, and angina are necessary to reduce the risks of random errors.

Following the publication of our review, we received an invitation from Heart to submit a summary of the review. This summary was published in 2018 [176].

# DISCUSSION

The crux of evidence-based medicine lies in randomised clinical trials [1], with systematic reviews of these trials regarded as the highest level of evidence assessing the effects of healthcare interventions [2]. There is a plethora of methodologies for analysing the results of systematic reviews and randomised clinical trials, with the ongoing development of new methods and novel applications of existing ones. However, certain methodologies are evidently more pivotal than others, presenting a challenge to researchers in selecting, combining, integrating and applying individual methodologies. This thesis explores the systemisation of crucial methodologies in randomised clinical trials and systematic reviews across various medical fields.

This thesis is divided into two parts.

The initial part of the thesis comprises eight theoretical papers aimed at systematising pivotal methodologies in randomised clinical trials and systematic reviews. The selection of each paper's specific topic was meticulously guided by experiences in conducting trials and systematic reviews at The Copenhagen Trial Unit during the last two decades. Central to this thesis are two papers that offer a comprehensive guide on determining whether statistical and clinical significance thresholds have been met in systematic reviews and randomised clinical trials [14, 17]. These two papers have been cited in over 290 distinct studies, trials, and systematic reviews ( https://scholar.google.dk/citations?user=gM2bze8AAAAJ&hl=da ). This thesis also includes six other papers that systematise additional key methodologies: 'When and how should multiple imputation be used for handling missing data in randomised clinical trials', 'Taking into account risks of random errors when analysing multiple outcomes in systematic reviews', 'Assessing assumptions for statistical analyses in randomised clinical trials', 'Assessment of assumptions of statistical analysis methods in randomised clinical trials: the what, and how', 'Power estimations for non-primary outcomes in randomised clinical trials', and 'Count data analysis in randomised clinical trials'.

The second section of this thesis consists of three systematic reviews: 'Direct-acting antivirals for chronic hepatitis C', 'Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder', and 'Drug-eluting stents versus bare-metal stents for acute coronary syndrome' [85, 117, 175]. These three reviews [85, 117, 175] are based on the methodologies described in the first section of this thesis. Two additional papers [100, 146] are included in this thesis summarising the key findings of the systematic reviews to clinicians and patients, including recommendations on how patients should be treated.

There may be debates suggesting that methodological 'cookbooks' are not credible scientific resources given that every randomised clinical trial or systematic review is distinct, as is every methodological issue within them. However, there are several reasons why these assertions may be debatable. First, numerous trials and systematic reviews based on the described methodologies have been published (this thesis incorporates just a tiny portion of these papers). These published works prove that applying the outlined combination of methodologies across research in various medical domains is feasible. Second, over the past two decades, the Copenhagen Trial Unit conducted research (both trials and reviews) spanning all medical fields. Throughout these years, it has been a consistent observation that the methodological challenges faced when performing clinical intervention research are fundamentally alike within a specific medical field and across different fields. Thus, it appears reasonable that analogous basic methodologies can be employed in research across medical disciplines. Third, it is difficult to improve and replicate a method if it is unclear what the methodology consists of. One of the founders of the modern scientific practice Ronald Fisher wrote in his 1935 book The Design of Experiments: “We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results” [177], and the philosopher Karl Popper wrote in his famous 1934 book The Logic of Scientific Discovery that “Non-reproducible single occurrences are of no significance to science” [178]. These declarations by two modern science trailblazers illustrate that research result reproducibility is a crucial prerequisite for research and science [177, 178]. The methodologies outlined in this thesis allow trialists and review authors to use a predetermined methodology that, if consistently used, will enhance the reproducibility and validity of research outcomes. Additionally, we have demonstrated in studies across medical fields that our systematised methodologies can be practically implemented. With a guide to select from hundreds of diverse methods and methodologies and some form of methodology systematisation, replication and improvement of the methodology appear easier. Thus, systematising the methodology applied within randomised clinical trials and systematic reviews is crucial for the scientific validity of clinical intervention research.

The strengths and limitations of each paper included in this thesis have been described previously (see ‘What we showed’ for each article summary). In general, the methodology used in the current thesis has three main limitations. First, the eight-step procedure for systematic reviews [17], the five-step procedure for randomised clinical trials [14], and the remaining methodologies [45, 46, 50, 61, 69, 80] outlined in this thesis are all founded on well-established methodologies. However, we are missing both simulation studies and comparative empirical studies that compare the utilisation of these suggested methods with 'standard practice'. To affirm that the quality of research improves when these methodologies (outlined in the initial section of this thesis) are utilised, simulation studies and empirical studies need to be carried out. Second, the uncertainty associated with quantifying anticipated intervention effects is a limitation of both the proposed eight-step and the five-step procedure, as it is in Bayesian analyses. However, estimating an anticipated intervention effect is compulsory when calculating a sample size in a trial, determining a required information size in a meta-analysis, or calculating the Bayes factor [100, 146]. We recommend estimating the anticipated intervention effect based on prior randomised clinical trials [14, 17] to ensure estimations are as optimal and objective as possible. If this is done, the results from the trials used to estimate the anticipated intervention effects will likely be reused in the comprehensive review analysis when conducting Trial Sequential Analysis. Thus, the required information size will take the form of an adaptive estimation [14, 17]. Therefore, the risk of type I error will increase [14, 17]. Due to this risk of circular reasoning, it is essential to adjust the required information size by applying a penalty for the weight of data from prior systematic reviews (or trials) [14, 17]. When the anticipated intervention effect is grounded in a predefined 'minimal important difference', there is no requirement for further adjustments [14, 17]. Although the need for additional adjustments is theoretically recognised, this would complicate the review analysis as the methodology would vary based on how the anticipated intervention effects are determined. Moreover, our advice already significantly tightens the significance thresholds in systematic reviews. If these thresholds are overly strict, there is a risk of 'throwing the baby out with the bath water' [14, 17]. Finally, if the methodology outlined in this thesis is systematically used, both statistical and clinical significance thresholds will become more stringent. More randomised participants will be needed before an intervention can be deemed 'evidence-based'. These rigorous thresholds may be criticised as excessively conservative and impractical due to unattainable sample sizes. However, such criticisms are not rooted in scientific doubt about the methodology's validity but rather in unsupported pragmatic arguments. To ensure patients receive treatments that offer more benefits than harm, we must employ valid methods and continuously strive to enhance and fine-tune these methodologies, even if it means larger sample sizes and more interventions with ambiguous effects.

In recent years, the scientific community has concentrated on the risks of systematic errors ('bias') in randomised clinical trials. Numerous meta-epidemiological studies have indicated a significant risk of overestimating benefits and underestimating harms when a trial has a high bias risk [58, 91-99]. The methodologies described in this thesis are motivated by the belief that the validity of clinical intervention research outcomes will improve if more consideration is given to the risks of random errors ('play of chance'). If we do not account for random error risks, the validity of research results may be jeopardised by multiple outcome comparisons, inadequate sample sizes, and insufficient information sizes.

The methodologies outlined in this dissertation are only a fraction of the numerous essential methodologies that require systematic documentation and explanation. For instance, our team is looking into the possibility that comprehensive Bayesian statistics may occasionally enhance the credibility of research findings compared to conventional frequentist statistics. We are also organising a consensus study to pinpoint the most effective method to account for 'centre' in multi-centre trials, while other teams are reportedly planning other significant methodological studies. This dissertation aims to add essential strategies; however, it only covers some methods, and there might be other methodologies that are just as, or more, valid than the ones described here.

This dissertation incorporates three systematic reviews using the eight-step procedure [85, 117, 177], alongside two articles discussing the clinical impact of the research findings derived from the eight-step procedure [100, 146]. These articles have achieved a certain level of success. The review on the effects of DAAs [85] was chosen as one of the top five notable papers at the International Liver Congress 2018 (EASL ILC 2018), and the British Medical Journal (BMJ) also invited us to compose an article about 'Uncertainties' drawing from our review [100]. The systematic review evaluating the impact of SSRIs on major depressive disorder has obtained the third-highest Almetric score among all publications from BMC Psychiatry (https://www.altmetric.com/details/16234955#score) [117]; likewise, 'Should antidepressants be used for major depressive disorder' [146] has acquired the fourth highest Almetric score among all articles from BMJ Evidence-based Medicine and was the most accessed article in 2019 (https://ebm.bmj.com/). Last, after the publication of the Cochrane review assessing the effects of drug-eluting stents versus bare metal stents [175], we were invited by BMJ Heart to pen a summary [176].

All the strategies discussed in this dissertation for analysing randomised clinical trials have been applied in two imminent randomised clinical trials. These include the largest-ever trial randomising cardiac arrest participants (the TTM2 trial) [179] and the most extensive placebo-controlled trial ever conducted to evaluate the effects of metformin [180]. Nonetheless, numerous trials have utilised the five-step procedure to analyse the outcomes of randomised clinical trials ( https://scholar.google.dk/scholar?oi=bibs&hl=da&cites=15733739536911770665&as_sdt=5 ).

Numerous papers discussing methodological concerns in randomised clinical trials and systematic reviews are published every week. New methodologies are constantly being developed, and existing ones are being utilised innovatively. However, some methodologies are evidently more critical than others, creating a potential dilemma for researchers when selecting, applying, and integrating different methodologies. This dissertation proposes a particular combination of methodologies and provides a practical guide on implementing this combination and examples of practical applications. It appears necessary for trialists and systematic review authors to employ a systematised methodology encompassing the most crucial methodological and statistical methods while considering both systematic errors ('bias') and random errors ('play of chance'). Adherence to a systematised methodology could enhance the validity of randomised clinical trials and systematic reviews with meta-analysis.

# DANISH SUMMARY

Det randomiserede kliniske forsøg er afgørende for evidensbaseret medicin, og den systematiske litteraturoversigt bør være øverst i evidenshierarkiet. Når man analyserer resultater fra randomiserede kliniske forsøg og systematiske litteraturoversigter, er der utallige metoder til rådighed – nye metoder bliver konstant udviklet, og kendte metoder anvendes på nye måder. Nogle metodologier synes mere vigtige end andre, og det kan ofte være uklart for forskere hvilke metoder der skal vælges, hvordan de skal anvendes og hvordan de bør kombineres.

Første del af denne afhandling indeholder otte teoretiske artikler som beskriver en systematisk fremgangsmåde ved analyse af systematiske litteraturoversigter samt randomiserede kliniske forsøg. Emnerne er valgt baseret på den erfaring med randomiserede kliniske forsøg og systematiske litteraturoversigter som Copenhagen Trial Unit har opbygget over de seneste årtier. Omdrejningspunktet i denne afhandling er de to 'Threshold' artikler, der beskriver en systematisk tilgang til at undersøge om grænsen for statistisk- og klinisk signifikans er nået.

Den anden del af denne afhandling indeholder tre systematiske litteraturgennemgange, hvor den foreslåede systematiske metode er blevet anvendt i praksis, samt to artikler der forsøger at formidle og fortolke forskningsresultaterne til klinikere og patienter.

Denne afhandling foreslår en specifik kombination af metodologier inklusiv en praktisk guide om hvordan denne metodologi kan anvendes systematisk. Hvis den systematiserede metodologi beskrevet i denne afhandling benyttes, vil gyldigheden af resultater fra randomiserede kliniske forsøg og systematiske litteraturanmeldelser øges.

# REFERENCES

1. Jakobsen JC, Gluud C: The necessity of randomized clinical trials. Br J Med Res 2013, 3(4):1453-1468.

2. Garattini S, Jakobsen JC, Wetterslev J, Bertele' V, Banzi R, Rath A, Neugebauer EAM, Laville M, Masson Y, Hivert V et al: Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. European Journal of Internal Medicine 2016, Jul;32:13-21. doi: 10.1016/j.ejim.2016.03.020.

3. Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K, Neumann I, Carrasco-Labra A, Agoritsas T, Hatala R et al: How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. Jama 2014, 312(2):171-179.

4. Sun X, Guyatt G: Meta-analysis of randomized trials for health care interventions: one for all? Journal of evidence-based medicine 2009, 2(1):53-56.

5. Borzak S, Ridker PM: Discordance between meta-analyses and large-scale randomized, controlled trials. Examples from the management of acute myocardial infarction. Ann Intern Med 1995, 123(11):873-877.

6. Hennekens CH, DeMets D: The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. Jama 2009, 302(21):2361-2362.

7. Stegenga J: Is meta-analysis the platinum standard of evidence? Studies in history and philosophy of biological and biomedical sciences 2011, 42(4):497-507.

8. Borm GF, Lemmers O, Fransen J, Donders R: The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol 2009, 62(7):711-715.e711.

9. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG: Evaluating non-randomised intervention studies. Health technology assessment (Winchester, England) 2003, 7(27):iii-x, 1-173.

10. Inthout J, Ioannidis JP, Borm GF: Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat Methods Med Res 2012.

11. Ioannidis J: Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005, 294(2):218-228.

12. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J: Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001, 286(7):821-830.

13. Higgins JPT, Green S: The Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Cochrane Collaboration 2011, Available from www.cochrane-handbook.org (latest access December 2016).

14. Jakobsen JC, Gluud C, Winkel P, Lange T, Wetterslev J: The thresholds for statistical and clinical significance - a five-step procedure for evaluation of intervention effects in randomised clinical trials. BMC medical research methodology 2014, 14(34).

15. Keus F, Wetterslev J, Gluud C, van Laarhoven CJ: Evidence at a glance: error matrix approach for overviewing available evidence. BMC medical research methodology 2010, 10:90.

16. Clarke M, Brice A, Chalmers I: Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PloS one 2014, 9(7):e102670.

17. Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C: Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. BMC medical research methodology 2014, 14:120.

18. Agoritsas T, Guyatt GH: Evidence-based medicine 20 years on: a view from the inside. Can J Neurol Sci 2013, 40(4):448-449.

19. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H et al: GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011, 64(4):383-394.

20. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, Atkins D, Kunz R, Montori V, Jaeschke R et al: GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol 2013, 66(2):151-157.

21. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G et al: GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011, 64(12):1283-1293.

22. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, Johnston BC, Karanicolas P, Akl EA, Vist G et al: GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013, 66(2):173-183.

23. Bjelakovic G, Gluud Lise L, Nikolova D, Whitfield K, Wetterslev J, Simonetti Rosa G, Bjelakovic M, Gluud C: Vitamin D supplementation for prevention of mortality in adults. In: Cochrane Database of Systematic Reviews. John Wiley & Sons, Ltd; 2014.

24. Wetterslev J, Thorlund K, Brok J, Gluud C: Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008, 61(1):64-75.

25. Wetterslev J, Thorlund K, Brok J, Gluud C: Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC medical research methodology 2009, 9:86.

26. Bauer P, Kohne K: Evaluation of experiments with adaptive interim analyses. Biometrics 1994, 50(4):1029-1041.

27. Roloff V, Higgins JP, Sutton AJ: Planning future studies based on the conditional power of a meta-analysis. Statistics in medicine 2013, 32(1):11-24.

28. Brok J, Thorlund K, Gluud C, Wetterslev J: Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analysis. J Clin Epidemiol 2008(61):763-769.

29. Pereira TV, Horwitz RI, Ioannidis JP: Empirical evaluation of very large treatment effects of medical interventions. JAMA 2012, 308:1676-1684.

30. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, Gluud LL, Als-Nielsen B, Gluud C: Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009, 38(1):276-286.

31. Ioannidis JPA: Why most published research findings are false. PLoS medicine 2005, 2(8):e124.

32. Johnson VE: Revised standards for statistical evidence. Proc Natl Acad Sci U S A 2013, 110(48):19313-19317.

33. Fisher R: Statistical methods and scientific induction. J R Stat Soc Ser B 1955, 17(1):69-78.

34. Gigerenzer G: Mindless statistics. J Socio Econ 2004, 33(5):587-606.

35. Hald A: "A history of mathematical statistics from 1750 to 1930", New-York-Chicheste-Weinhei-Brisban-Singapor-Toronto: John Wiley & Sons; 1998.

36. Oliveri RS, Gluud C, Wille-Jørgensen PA: Hospital doctors' self-rated skills in and use of evidence-based medicine - a questionnaire survey. J Eval Clin Pract 2004, 10(2):219-226.

37. Goodman S: A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008, 45:135-140.

38. Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, Heels-Ansdell D, Walter SD, Guyatt GH, Flynn DN et al: Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010, 303:1180-1187.

39. Groenwold RHH, Moons KGM, Vandenbroucke JP: Randomized trials with missing outcome data: how to analyze and what to report. CMAJ : Canadian Medical Association Journal 2014, 186(15):1153-1157.

40. Nguyen TL, Collins GS, Lamy A, Devereaux PJ, Daures JP, Landais P, Le Manach Y: Simple randomization did not protect against bias in smaller trials. J Clin Epidemiol 2017, 84:105-113.

41. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA et al: The prevention and treatment of missing data in clinical trials. The New England journal of medicine 2012, 367(14):1355-1360.

42. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clinical research ed) 2009, 338.

43. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P: Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med 2013, 86(3):343-358.

44. Morris TP, Kahan BC, White IR: Choosing sensitivity analyses for randomised trials: principles. BMC medical research methodology 2014, 14:11.

45. Jakobsen JC, Gluud C, Wetterslev J, Winkel P: When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC medical research methodology 2017, 17(1):162.

46. Jakobsen JC, Wetterslev J, Lange T, Gluud C: Viewpoint: taking into account risks of random errors when analysing multiple outcomes in systematic reviews [editorial]. Cochrane Database of Syst Rev 2016(3).

47. Imberger G, Vejlby AD, Hansen SB, Møller AM, Wetterslev J: Statistical multiplicity in systematic reviews of anaesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews. PloS one 2011, 6:e28422.

48. Korn EL, Li MC, McShane LM, Simon R: An investigation of two multivariate permutation methods for controlling the false discovery proportion. Statistics in medicine 2007, 26(24):4428-4440.

49. Altman DG, Bland JM: How to obtain the confidence interval from a P value. BMJ (Clinical research ed) 2011, 343:d2090.

50. Nielsen EE, Norskov AK, Lange T, Thabane L, Wetterslev J, Beyersmann J, de Una-Alvarez J, Torri V, Billot L, Putter H et al: Assessing assumptions for statistical analyses in randomised clinical trials. BMJ Evid Based Med 2019.

51. Jakobsen JC, Gluud C, Winkel P, Lange T, Wetterslev J: The thresholds for statistical and clinical significance - a five-step procedure for evaluation of intervention effects in randomised clinical trials. BMC Med Res Methodol 2014, 14(1):34.

52. Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C: Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. BMC Med Res Methodol 2014, 14(1):120.

53. Piantadosi S: Clinical Trials: A Methodologic Perspective, Second edn: Wiley; 2005.

54. Pocock SJ: Randomised clinical trials. Br Med J 1977, 1(6077):1661.

55. Enhancing the QUAlity and Transparency Of health Research

56. Garattini S, Jakobsen JC, Wetterslev J, Bertele' V, Banzi R, Rath A, Neugebauer EAM, Laville M, Masson Y, Hivert V et al: Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. Eur J Intern Med 2016, 32:13-31.

57. Schulz KF, Altman Dg Fau - Moher D, Moher D: CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010, 152(11):726-732.

58. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, Als-Nielsen B, Balk EM, Gluud C, Gluud LL et al: Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012, 157(6):429-438.

59. Sawilowsky SS: Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney U test for shift in location parameter. J Mod Appl Stat Methods 2005, 4(2):598-600.

60. Nielsen EE, Norskov AK, Lange T, Thabane L, Wetterslev J, Beyersmann J, de Una-Alvarez J, Torri V, Billot L, Putter H et al: Assessing assumptions for statistical analyses in randomised clinical trials. BMJ Evid Based Med 2019, 24(5):185-189.

61. Norskov AK LT, Nielsen EE, Thabane L, Wetterslev J, Beyersmann J, de Una-Alvarez J, Torri V, Billot L, Putter H, Winkel P, Gluud C, Jakobsen JC: Assessment of Assumptions of Statistical Analysis Methods in Randomised Clinical Trials: the What, and How. BMJ Evid Based Med 2019 (accepted for publication) 2020.

62. Kahan BC, Morris TP: Improper analysis of trials randomised using stratified blocks or minimisation. Statistics in medicine 2011, 31:328-340.

63. Kahan BC, Morris TP: Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. BMJ (Clinical research ed) 2012, 345:e5840.

64. Anderson TW: Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes. Ann Math Stat 1952, 23(2):193-212.

65. Shapiro SS, M.B W: An analysis of variance test for normality (complete samples). Biometrika 1962, 52(3-4):591-611.

66. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin JA et al: SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. Ann Intern Med 2013, 158(3):200-207.

67. Schoenfeld D: Partial Residuals for The Proportional HAzards Regression Model. Biometrika 1982, 69(1):239-241.

68. Brown BB: Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts Santa Monica, CA: RAND Corporation, 1968. https://www.rand.org/pubs/papers/P3925.html. Also available in print form.

69. Jakobsen JC, Ovesen C, Winkel P, Hilden J, Gluud C, Wetterslev J: Power estimations for non-primary outcomes in randomised clinical trials. BMJ Open 2019, 9(6):e027092.

70. Schulz KF, Altman DG, Moher D: CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Int Med 2010, 152(11):726-732.

71. European Medicines Agency: Guideline on multiplicity issues in clinical trials. Committee for Human Medicinal Products (CHMP) EMA/CHMP/44762/2017 2016.

72. Lan GKK, DeMets DL: Discrete sequential boundaries for clinical trials. Biometrika 1983, 70(3):659-663.

73. Zhang Y, Zhang S, Thabane L, Furukawa TA, Johnston BC, Guyatt GH: Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. J Clin Epidemiol 2015, 68(8):888-894.

74. StataCorp.: Stata: Release 15. Statistical Software. College Station, TX: StataCorp LP. 2017.

75. Winkel P, Bath PM, Gluud C, Lindschou J, van der Worp HB, Macleod MR, Szabo I, Durand-Zaleski I, Schwab S, Euro HYPti: Statistical analysis plan for the EuroHYP-1 trial: European multicentre, randomised, phase III clinical trial of the therapeutic hypothermia plus best medical treatment versus best medical treatment alone for acute ischaemic stroke. Trials 2017, 18(1):573.

76. Epidural Position Trial Collaborative Group: Upright versus lying down position in second stage of labour in nulliparous women with low dose epidural: BUMPES randomised controlled trial. BMJ (Clinical research ed) 2017, 359:j4471.

77. Kirkham JJ, Gorst S, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, Moher D, Schmitt J, Tugwell P et al: Core Outcome Set-STAndards for Reporting: The COS-STAR Statement. PLoS medicine 2016, 13(10):e1002148.

78. Karazsia BT, van Dulmen MH: Regression models for count data: illustrations using longitudinal predictors of childhood injury. Journal of pediatric psychology 2008, 33(10):1076-1084.

79. Agresti A: An Introduction to Categorical Data Analysis (Wiley Series in Probability and Statistics): Wiley-Interscience; 2007.

80. Jakobsen JC, Tamborrino M, Winkel P, Haase N, Perner A, Wetterslev J, Gluud C: Count data analysis in randomised clinical trials. J Biomet Biostat 2015, 6(1):227.

81. Dmitrienko A, Ajit C. Tamhane AC, Bretz F: Multiple testing problems in pharmaceutical statistics (Chapman & Hall/CRC Biostatistics Series): Chapman and Hall/CRC; 2009.

82. The European Agency for the Evaluation of Medical Products: Points to consider on multiplicity issues in clinical trials. 2002, CPMP/EWP/908/99.

83. The Nordic Trial Alliance Working Group 6: Report on transperancy and registration in clinical research in the Nordic countries. Project WP 6: Transparency and registration 2015.

84. WHO: Global hepatitis C report. World Health Association http://wwwwhoint/mediacentre/factsheets/fs164/en/ 2017.

85. Jakobsen JC, Nielsen EE, Feinberg J, Katakam KK, Fobian K, Hauser G, Poropat G, Djurisic S, Weiss KH, Bjelakovic M et al: Direct-acting antivirals for chronic hepatitis C. Cochrane Database of Systematic Reviews 2017(9).

86. Chopra S: Clinical manifestations and natural history of chronic hepatitis C virus infection. UpToDate, Inc available at http://www.uptodatecom/index 2017, Accessed 15th of February 2018.

87. Pearlman BL, Traub N: Sustained virologic response to antiviral therapy for chronic hepatitis C virus infection: a cure and so much more. Clin Infect Dis 2011, 52(7):889-900.

88. Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 2015, 62(3):932-954.

89. EASL Recommendations on Treatment of Hepatitis C 2016. J Hepatol 2017, 66(1):153-194.

90. WHO: Guidelines for the Screening Care and Treatment of Persons with Chronic Hepatitis C Infection: Updated Version WHO Guidelines Approved by the Guidelines Review Committee 2016, Available at: http://www.who.int/hepatitis/publications/hepatitis-c-guidelines-2016/en/.

91. Ebrahim S, Bance S, Athale A, Malachowski C, Ioannidis JP: Meta-analyses with industry involvement are massively published and report no caveats for antidepressants. J Clin Epidemiol 2016, 70:155-163.

92. Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S: Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol 2014, 43(4):1272-1283.

93. Hróbjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S: Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ (Clinical research ed) 2012, 344:e1119.

94. Hróbjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S: Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 2013, 185(4):E201-211.

95. Lundh A, Lexchin J, Mintzes B, Scholl JB, Bero L: Industry sponsorship and research outcome. Cochrane Database Syst Rev 2017, Art. No.: MR000033. DOI: 10.1002/14651858.MR000033.pub3.(2):MR000033.

96. Nüesch E, Trelle S, Reichenbach S, Rutjes AWS, Bürgi E, Scherer M, Altman DG, Jüni P: The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ (Clinical research ed) 2009, 339.

97. Savovic J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, Sterne JAC: Association Between Risk-of-Bias Assessments and Results of Randomized Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study. American journal of epidemiology 2018, 187(5):1113-1122.

98. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR: Empirical assessment of effect of publication bias on meta-analyses. BMJ (Clinical research ed) 2000, 320(7249):1574-1577.

99. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA: Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ (Clinical research ed) 2008, 336(7644):601-605.

100. Jakobsen JC, Nielsen EE, Koretz RL, Gluud C: Do direct acting antivirals cure chronic hepatitis C? BMJ (Clinical research ed) 2018, 361.

101. Morgan RL, Baack B, Smith BD, Yartel A, Pitasi M, Falck-Ytter Y: Eradication of hepatitis C virus infection and the development of hepatocellular carcinoma: a meta-analysis of observational studies. Ann Intern Med 2013, 158(5 Pt 1):329-337.

102. van der Meer AJ, Veldt BJ, Feld JJ, Wedemeyer H, Dufour JF, Lammert F, Duarte-Rojo A, Heathcote EJ, Manns MP, Kuske L et al: Association between sustained virological response and all-cause mortality among patients with chronic hepatitis C and advanced hepatic fibrosis. Jama 2012, 308(24):2584-2593.

103. Veldt BJ, Heathcote EJ, Wedemeyer H, Reichen J, Hofmann WP, Zeuzem S, Manns MP, Hansen BE, Schalm SW, Janssen HL: Sustained virologic response and clinical outcomes in patients with chronic hepatitis C and advanced fibrosis. Ann Intern Med 2007, 147(10):677-684.

104. Koretz RL, Lin KW, Ioannidis JP, Lenzer J: Is widespread screening for hepatitis C justified? BMJ (Clinical research ed) 2015, 350:g7809 doi: https://doi.org/7810.1136/bmj.g7809.

105. Thorlund K, Walter SD, Johnston BC, Furukawa TA, Guyatt GH: Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. Research synthesis methods 2011, 2(3):188-203.

106. Chopra S, Pockros PJ: Overview of the management of chronic hepatitis C virus infection. UpToDate, Inc available at http://wwwuptodatecom/index 2017, Accessed 15th of February 2018.

107. Tsoulfas G, Goulis I, Giakoustidis D, Akriviadis E, Agorastou P, Imvrios G, Papanikolaou V: Hepatitis C and liver transplantation. Hippokratia 2009, 13(4):211-215.

108. Gualano MR, Bert F, Mannocci A, La Torre G, Zeppegno P, Siliquini R: Consumption of Antidepressants in Italy: Recent Trends and Their Significance for Public Health. Psychiatric services (Washington, DC) 2014, 65(10):1226-1231.

109. Wise J: GPs in England prescribed 2.7 million extra antidepressants during 2012 recession. BMJ (Clinical research ed) 2014, 348:g3607.

110. Gibbons RD, Hur K, Brown CH, Davis JM, Mann JJ: Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. Arch Gen Psychiatry 2012, 69(6):572-579.

111. Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J: Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 2010, 303(1):47-53.

112. Khan A, Leventhal RM, Khan SR, Brown WA: Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. J Clin Psychopharmacol 2002, 22(1):40-45.

113. Undurraga J, Baldessarini RJ: Randomized, placebo-controlled trials of antidepressants for acute major depression: thirty-year meta-analytic review. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 2012, 37(4):851-864.

114. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R: Selective publication of antidepressant trials and its influence on apparent efficacy. The New England journal of medicine 2008, 358(3):252-260.

115. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT: Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS Medicine 2008, 5(2):e45.

116. Wilson K, Mottram PG, Sivananthan A, Nightingale A: Antidepressants versus placebo for the depressed elderly. Cochrane Database Syst Rev 2001.

117. Jakobsen JC, Katakam KK, Schou A, Hellmuth SG, Stallknecht SE, Leth-Møller K, Iversen M, Banke MB, Petersen IJ, Klingenberg SL et al: Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder. A systematic review with meta-analysis and Trial Sequential Analysis. BMC Psychiatr 2017, 17(1):58.

118. Broich K: Committee for Medicinal Products for Human Use (CHMP) assessment on efficacy of antidepressants. Eur Neuropsychopharmacol 2009, 19(5):305-308.

119. Hunt SM, McKenna SP: The QLDS: a scale for the measurement of quality of life in depression. Health Policy 1992, 22(3):307-319.

120. Wetterslev J, Jakobsen JC, Gluud C: Forsøgssekventielle metaanalyser i systematiske oversigter. Bibl Laeger 2015, 207(2):124-151.

121. National Institute for Clinical Excellence: Depression: Mangement of Depression in Primary and Secundary Care. London, England: National Institute for Clinical Excellence 2004:640.

122. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT: Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS medicine 2008, 5(2):e45.

123. Rimer J, Dwan K, Lawlor Debbie A, Greig Carolyn A, McMurdo M, Morley W, Mead Gillian E: Exercise for depression. Cochrane database of systematic reviews (Online) 2012(7):CD004366.pub004366.

124. Higgins JPT, Green S: The Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Cochrane Collaboration 2011, Available from www.cochrane-handbook.org.

125. Leucht S, Fennema H, Engel R, Kaspers-Janssen M, Lepping P, Szegedi A: What does the HAMD mean? J Affect Disord 2013, 148(2-3):243-248.

126. Moncrieff J, Kirsch I: Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. Contemporary Clinical Trials 2015, 43:60-62.

127. Furukawa TA, Cipriani A, Atkinson LZ, Leucht S, Ogawa Y, Takeshima N, Hayasaka Y, Chaimani A, Salanti G: Placebo response rates in antidepressant trials: a systematic review of published and unpublished double-blind randomised controlled studies. The Lancet Psychiatry, 3(11):1059-1066.

128. Hrobjartsson A, Gotzsche PC: Placebo interventions for all clinical conditions. Cochrane Database Syst Rev 2010(1):CD003974.

129. Andrews P, Thomson JA, Amstadter A, Neale M: Primum Non Nocere: An Evolutionary Analysis of Whether Antidepressants Do More Harm than Good. Frontiers in Psychology 2012, 3(117).

130. Gøtzsche PC: Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare: Radcliffe Medical Press ltd, London, N1 7LH United Kingdom; 2013.

131. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H et al: Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 2009, 373(9665):746-758.

132. World Health Orginization (WHO): Depression (fact sheet). http://wwwwhoint/mediacentre/factsheets/fs369/en/ assessed April 2018 2018.

133. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC: Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. Scientific reports 2018, 8(1):2861-2861.

134. Hasin DS, Sarvet AL, Meyers JL, Saha TD, Ruan WJ, Stohl M, Grant BF: Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States. JAMA Psychiatry 2018, 75(4):336-346.

135. OECD (2017): “Antidepressant drugs consumption, 2000 and 2015 (or nearest year)”, in Pharmaceutical sector, OECD Publishing, Paris, http://dx.doi.org/10.1787/health_glance-2017-graph181-en. 2017.

136. Pratt LA, Brody DJ, Gu Q: Antidepressant use among persons aged 12 and over: United States, 2011–2014. NCHS data brief, no 283. Hyattsville, MD: National Center for Health Statistics. 2017.

137. Cuijpers P, van Straten A, Warmerdam L, Andersson G: Psychotherapy versus the combination of psychotherapy and pharmacotherapy in the treatment of depression: a meta-analysis. Depress Anxiety 2009, 26(3):279-288.

138. Barbateskovic M, Marker S, Jakobsen JC, Krag M, Granholm A, Anthon CT, Perner A, Wetterslev J, Møller MH: Stress ulcer prophylaxis in adult intensive care unit patients – a protocol for a systematic review. Acta Anaesthesiologica Scandinavica 2018(0).

139. Bauer M, Severus E, Moller HJ, Young AH, Disorders WTFoUD: Pharmacological treatment of unipolar depressive disorders: summary of WFSBP guidelines. Int J Psychiatry Clin Pract 2017, 21(3):166-176.

140. Cleare A, Pariante CM, Young AH, Anderson IM, Christmas D, Cowen PJ, Dickens C, Ferrier IN, Geddes J, Gilbody S et al: Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. J Psychopharmacol 2015, 29(5):459-525.

141. Gelenberg AJ, Freeman MP, Markowitz JC, Rosenbaum JF, Thase M, Trivedi MH, Van Rhoads RS: Practice Guideline for the Treatment of Patients With Major Depressive Disorder. In., 3 edn: American Psychiatric Association; 2010.

142. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, Hasnain M, Jollant F, Levitt AJ, MacQueen GM et al: Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments. Canadian journal of psychiatry Revue canadienne de psychiatrie 2016, 61(9):540-560.

143. Malhi GS, Bassett D, Boyce P, Bryant R, Fitzgerald PB, Fritz K, Hopwood M, Lyndon B, Mulder R, Murray G et al: Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for mood disorders. The Australian and New Zealand journal of psychiatry 2015, 49(12):1087-1206.

144. Hirsch M, Birnbaum RJ: Switching antidepressant medications in adults. UpToDate, Inc available at http://wwwuptodatecom/index 2017.

145. Marcus SC, Olfson M: National trends in the treatment for depression from 1998 to 2007. Arch Gen Psychiatry 2010, 67(12):1265-1273.

146. Jakobsen JC, Gluud C, Kirsch I: Should antidepressants be used for major depressive disorder? BMJ Evid Based Med 2019.

147. National Institute for Clinical Excellence. Depression: management of depression. Primary and secondary care Clinical practice guideline No 23 London: NICE, 2004 wwwniceorguk/pageaspx?o=235213.

148. Crosby RD, Kolotkin RL, Williams GR: Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003, 56(5):395-407.

149. Moncrieff J, Kirsch I: Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. Cont Clin Trials 2015, 43:60-62.

150. Altman DG, Royston P: The cost of dichotomising continuous variables. BMJ (Clinical research ed) 2006, 332(7549):1080.

151. Kirsch I, Moncrieff J: Clinical trials and the response rate illusion. Contemp Clin Trials 2007, 28(4):348-351.

152. Tennant P: Antidepressant Benefits: Misinferance from ordinal scales? BMJ (Clinical research ed) 2008, 336:466doi: https://doi.org/10.1136/bmj.39503.656852.DB.

153. Bagby RM, Ryder AG, Schuller DR, Marshall MB: The Hamilton Depression Rating Scale: has the gold standard become a lead weight? American Journal of Psychiatry 2004, 161(12):2163-2177.

154. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ (Clinical research ed) 2009, 339:b2700.

155. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, Leucht S, Ruhe HG, Turner EH, Higgins JPT et al: Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. The Lancet 2018.

156. Wisniewski SR, Rush AJ, Nierenberg AA, Gaynes BN, Warden D, Luther JF, McGrath PJ, Lavori PW, Thase ME, Fava M et al: Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR*D report. The American journal of psychiatry 2009, 166(5):599-607.

157. Kirsch I, Huedo-Medina T, Pigott H, Johnson B: Do outcomes of clinical trials resemble those “real world” patients? A reanalysis of the STAR*D antidepressant data set.”. Psychology of Consciousness: Theory, Research, and Practice 2018.

158. Furukawa TA, Maruo K, Noma H, Tanaka S, Imai H, Shinohara K, Ikeda K, Yamawaki S, Levine SZ, Goldberg Y et al: Initial severity of major depression and efficacy of new generation antidepressants: individual participant data meta-analysis. Acta psychiatrica Scandinavica 2018, 137(6):450-458.

159. Rabinowitz J, Werbeloff N, Mandel FS, Menard F, Marangell L, Kapur S: Initial depression severity and response to antidepressants v. placebo: patient-level data analysis from 34 randomised controlled trials. The British journal of psychiatry : the journal of mental science 2016, 209(5):427-428.

160. Henssler J, Kurschus M, Franklin J, Bschor T, Baethge C: Long-Term Acute-Phase Treatment With Antidepressants, 8 Weeks and Beyond: A Systematic Review and Meta-Analysis of Randomized, Placebo-Controlled Trials. The Journal of clinical psychiatry 2018, 79(1).

161. Fava GA: Can long-term treatment with antidepressant drugs worsen the course of depression? The Journal of clinical psychiatry 2003, 64(2):123-133.

162. Ferguson JM: SSRI Antidepressant Medications: Adverse Effects and Tolerability. Prim Care Companion J Clin Psychiatry 2001, 3(1):22-27.

163. Farnsworth KD, Dinsmore WW: Persistent sexual dysfunction in genitourinary medicine clinic attendees induced by selective serotonin reuptake inhibitors. Int J STD AIDS 2009, 20(1):68-69.

164. Reefhuis J, Devine O, Friedman JM, Louik C, Honein MA, National Birth Defects Prevention S: Specific SSRIs and birth defects: Bayesian analysis to interpret new data in the context of previous reports. BMJ (Clinical research ed) 2015, 351:h3190.

165. Fava GA, Gatti A, Belaise C, Guidi J, Offidani E: Withdrawal Symptoms after Selective Serotonin Reuptake Inhibitor Discontinuation: A Systematic Review. Psychother Psychosom 2015, 84(2):72-81.

166. Davies J, Read J: A systematic review into the incidence, severity and duration of antidepressant withdrawal effects: Are guidelines evidence-based? Addictive Behaviors 2018.

167. Geddes JR, Carney SM, Davies C, Furukawa TA, Kupfer DJ, Frank E, Goodwin GM: Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. Lancet 2003, 361(9358):653-661.

168. Reid S, Barbui C: Long term treatment of depression with selective serotonin reuptake inhibitors and newer antidepressants. BMJ (Clinical research ed) 2010, 340:c1468.

169. National Institute for Health and Care Excellence (NICE). Depression in adults: recognition and management. Clinical guideline [CG90] Published date: October 2009 Last updated: April 2016.

170. American Psychiatric Association: Practice Guideline for the Treatment of Patients with Major Depressive Disorder, Third Edition, 2010. http://psychiatryonline.org/guidelines.aspx.

171. Parikh SV, Quilty LC, Ravitz P, Rosenbluth M, Pavlova B, Grigoriadis S, Velyvis V, Kennedy SH, Lam RW, MacQueen GM et al: Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 2. Psychological Treatments. Canadian journal of psychiatry Revue canadienne de psychiatrie 2016, 61(9):524-539.

172. Nichols M, Townsend N, Scarborough P, Rayner M: Cardiovascular disease in Europe 2014: epidemiological update. European heart journal 2014, 35(42):2929.

173. World Health Organization (WHO): Cardiovascular diseases. www.who.int/cardiovascular_diseases/en/ (accessed 12 January 2019).

174. Schmidt M, Jacobsen JB, Lash TL, Botker HE, Sorensen HT: 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity: a Danish nationwide cohort study. BMJ (Clinical research ed) 2012, 344:e356.

175. Feinberg J, Nielsen EE, Greenhalgh J, Hounsome J, Sethi NJ, Safi S, Gluud C, Jakobsen JC: Drug-eluting stents versus bare-metal stents for acute coronary syndrome. Cochrane Database of Systematic Reviews 2017(8).

176. Feinberg J, Nielsen EE, Gluud C, Jakobsen JC: Cochrane Corner: drug-eluting stents versus bare-metal stents for acute coronary syndrome. Heart (British Cardiac Society) 2018.

177. Fisher RA: The Design of Experiments: Macmillan; 1935.

178. Popper K: Logik der Forschung (The Logic of Scientific Discovery): Mohr Siebeck; 1934.

179. Jakobsen JC, Lange T, Ullén S, Dankiewicz J, Cronberg T, Lilja G, Bělohlávek J, Callaway C, Cariou A, Erlinge D et al: Detailed statistical analysis plan for the targeted hypothermia versus targeted normothermia after out-of-hospital cardiac arrest randomised clinical trial. Submitted to TRIALS JAN 2020.

180. Jakobsen JC, Madsbad S, Hemmingsen B, Ovesen C, Gluud C, Sneppen SB, Breum L, Hedetoft C, Krarup T, Lundby-Christensen L et al: Detailed statistical analysis plan for the outcomes quality of life, patient satisfaction, and cardiovascular outcomes of the randomised 2 x 3 factorial Copenhagen Insulin and Metformin Therapy (CIMT) trial. Submitted to TRIALS DEC 2019.

Du er den bedste far