entropy

MDPI

Article

A Note on Cherry-Picking in Meta-Analyses

Daisuke Yoneoka [1,*] and Bastian Rieck [2]

[1] Center for Surveillance, Immunization, and Epidemiologic Research, National Institute of Infectious Diseases, Tokyo 162-8640, Japan
[2] Institute of AI for Health, Helmholtz Munich, Technical University of Munich, 80333 Munich, Germany
* Correspondence: yoneoka@niid.go.jp; Tel.: +81-03-5285-1111

Abstract: We study selection bias in meta-analyses by assuming the presence of researchers (meta-analysts) who intentionally or unintentionally cherry-pick a subset of studies by defining arbitrary inclusion and/or exclusion criteria that will lead to their desired results. When the number of studies is sufficiently large, we theoretically show that a meta-analysts might falsely obtain (non)significant overall treatment effects, regardless of the actual effectiveness of a treatment. We analyze all theoretical findings based on extensive simulation experiments and practical clinical examples. Numerical evaluations demonstrate that the standard method for meta-analyses has the potential to be cherry-picked.

Keywords: meta-analysis; cherry-picking studies; selection bias; adversarial meta-analysis; inclusion/exclusion criteria

check for updates

Citation: Yoneoka, D.; Rieck, B. A Note on Cherry-Picking in Meta-Analyses. *Entropy* **2023**, *25*, 691. https://doi.org/10.3390/e25040691

Academic Editors: Shesh Nath Rai, Anil Rai and Samarendra Das

Received: 3 March 2023
Revised: 11 April 2023
Accepted: 18 April 2023
Published: 19 April 2023

1. Introduction

Meta-analysis is a methodology for evaluating the overall treatment effect by integrating the results of past clinical trials and is widely recognized as one of the research methods that underlie "Evidence Based Medicine" [1,2]. Generally, the methodology involves the integration of summary statistics, such as odds ratios or hazard ratios reported in published papers, by using appropriate statistical methods to estimate the average treatment effect [1–3]. In a meta-analysis, various biases that could affect the validity of the synthesized results have been widely studied, for example, (1) *publication bias*, whereby positive results are more likely than negative or null results to be published [4]; (2) *language bias*, whereby non-English studies tend to be excluded from meta-analyses [5]; (3) *time-lag bias*, whereby positive results tend to have longer time differences from trial completion to publication than negative or null results [6]; (4) *reporting bias*, whereby studies selectively report outcomes favoring their hypothesis [1]; (5) *outlier bias*, whereby a single or a few studies disproportionately influence the overall results of a meta-analysis [7]; (6) *categorization bias*, whereby studies use different categorization or stratification schemes to achieve the same outcome [8]; and (7) *covariate set bias*, whereby studies use different covariate sets in the regression model that share the same regression task across the studies [9]. In this study, we aim to focus on a new source of bias, the "cherry-picking" bias, in meta-analyses.

The simplest setup of a meta-analysis is to assume that there are K independent studies, each yielding an estimate y_i ($i = 1, \ldots, K$) of an underlying treatment effect parameter θ. The standard fixed-effect model is defined as

$$y_i \sim N(\theta, \sigma_i^2), \tag{1}$$

where σ_i^2 is the reported (known) within-study variance of the ith study. Under this fixed-effect model, the maximum likelihood estimate of θ is defined by the weighted average

$$\hat{\theta} = \frac{\sum_{i=1}^{K} w_i y_i}{\sum_{i=1}^{K} w_i}, \tag{2}$$

where the ith study is assigned the weight $w_i = 1/\sigma_i^2$. The corresponding standard normal test statistic is

$$T(\theta) = \sqrt{\sum_{i=1}^{K} w_i}(\hat{\theta} - \theta),$$

and the resulting confidence interval (CI) is

$$\{\theta : |T(\theta)| \leq z_\alpha\} = \left[\hat{\theta} - z_\alpha \left(\sum w_i\right)^{-1/2}, \hat{\theta} + z_\alpha \left(\sum w_i\right)^{-1/2}\right],$$

where $z_\alpha = \Phi^{-1}(1 - \alpha/2)$ is the standard normal percentage point for the coverage of $1 - \alpha$, and Φ is the cumulative distribution function of the standard normal distribution [10–13]. In practice, researchers are often interested in a hypothesis regarding whether a given treatment has no effect ($H_0 : \theta = 0$) or is beneficial ($H_1 : \theta > 0$). The one-sided p-value of the ith study is defined as

$$p_i = \Phi(-\sqrt{w_i} y_i). \tag{3}$$

Similarly, the p-value of Equation (2) is defined as

$$p_{meta} = \Phi\left(-\hat{\theta}\sqrt{\sum_{i=1}^{K} w_i}\right). \tag{4}$$

Equation (1) is based on the fixed-effect assumption that each study shares the same underlying effect θ. When heterogeneity between included studies is suspected, the random-effect model is fitted as

$$y_i \sim N(\theta, \sigma_i^2 + \tau^2), \tag{5}$$

where τ^2 is the between-study variance, which can be estimated from the data using standard methods, such as the method proposed in DerSimonian and Laird [3]. The same reasoning can be applied by replacing the weights w_i in Equation (2) with

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}. \tag{6}$$

Refer to the studies by [10–12] and Cooper et al. [13] for a detailed discussion of the various methods used for meta-analysis.

One of the most important stages of a meta-analysis is the specification of the inclusion and/or exclusion criteria, because the selection of studies for a literature review is known to influence the conclusions. One must carefully consider which studies to include or exclude from the review to obtain unbiased and fair conclusions. However, in reality, a significant number of meta-analyses are published without a protocol to define the inclusion and exclusion criteria before conducting the meta-analysis and systematic review. Furthermore, it is not common for papers to follow procedures such as stating inclusion and exclusion criteria in advance and adhering to them. For example, Page et al., (2016) examined the reporting completeness of Biomedical Research meta-analyses and found that only 16% of the included reviews had a publicly accessible protocol published before the review was conducted [14]. In addition, Tawfic et al., (2020) found that only 37.4% of researchers who are trying to conduct a meta-analysis agree that protocol registration prior to the main analysis should be mandatory [15].

Given a set of included studies, the conclusions obtained from the results of meta-analyses are frequently based on statistical tests and their associated p-values in practice. Ideally, a statistical test with a type 1 error rate of α should be used to control the ratio of

false findings at a ratio of (less than) α. However, inclusion and/or exclusion criteria can be misused by (sometimes malicious) meta-analysts (i.e., the authors of a meta-analysis who intentionally or unintentionally report false (non)significant overall effects, regardless of the actual treatment effect) to pick a subset of all studies that changes the result and sometimes leads to their desired conclusion. This practice is also known as cherry-picking, and it means that the resulting p-value no longer controls the ratio of false findings. Figures in Section 4 show practical examples. Reviewer selection bias is also known in the field of meta-analysis as the situation where reviewers (un)intentionally seek only a subset of existing studies that satisfy certain criteria, so the chosen subset does not reflect all available evidence [16]. The degree of bias in a synthesized result can depend on a selector's prior knowledge, research field, existing collaborators, and opinion regarding the research question of interest [17]. Other similar biases related to inclusion and/or exclusion criteria include the English language bias (whereby non-English studies are more likely to be excluded), the data availability bias (whereby only studies with individual patient data are included), and the database bias (whereby only studies published in journals indexed in popular databases such as Embase or Medline are included); see [17–19] for an overview of this topic. For instance, Ahmed et al. [17] investigated 31 meta-analyses and found that 29% of them suffered from a significant selection bias based on the use of selective or nonsystematic approaches for the identification of relevant studies. They concluded that biased synthesized results can lead to incorrect decisions by medical practitioners, which can harm patients because inefficient or ineffective treatments may be chosen. Such results can also mislead future research efforts [20]. However, although the selection bias has a similar impact on synthesized results to the publication bias, which has been widely studied in the field of meta-analysis, no attempts have been made to examine the selection bias from a statistical perspective. In this study, we demonstrate that it is possible to modify the results of a meta-analysis by changing the inclusion and/or exclusion criteria to select an arbitrary subset of studies, so that they support a biased conclusion, such as (i) the treatment of interest having a significant effect, despite there being no actual effect or (ii) the treatment having a nonsignificant effect, despite the presence of an actual effect. The reliability of a meta-analysis is decreased in the presence of such a selection bias. The goal of this study is to identify the possibility of cherry-picking.

The remainder of the article is organized as follows: In Section 2, we show theoretical guarantees on the chance of cherry-picking by meta-analysts who intentionally or unintentionally select the subset of studies. To demonstrate that conventional meta-analysis procedures have a significant cherry-picking effect, the results of extensive simulation studies are presented in Section 3, and two clinical datasets are examined in Section 4. Lastly, Section 5 presents a discussion and our conclusions.

2. Methods

We consider the simple fixed-effect meta-analysis settings defined in Equation (1). An extension for a random-effect model is described in Section 2.2 and later in the discussion section. We assume that there are K studies $\mathcal{D}_K = \{1, 2, \ldots, K\}$ collected via data extraction from several databases such as PubMed, Medline, and Embase. Each study is supposed to report an estimate y_i and corresponding variance σ_i^2 (or w_i, equivalently). Meta-analysts determine the inclusion and/or exclusion criteria to select a subset of S studies from all K studies found in the databases. This subset is denoted as $\mathcal{D}_S = \{1, 2, \ldots, S\} \subseteq \mathcal{D}_K$. Therefore, $\mathcal{D}_S$ may suffer from a selection bias. In this study, we assume that meta-analysts (intentionally or unintentionally) select studies $\mathcal{D}_S$ to (i) overstate the effect of the treatment of interest (Case 1), despite the treatment having no actual effect (i.e., $\theta = 0$), or (ii) understate the effect of the treatment (Case 2), despite the treatment having an actual effect (i.e., $\theta > 0$). Furthermore, we assume that meta-analysts use a statistical testing framework by defining the null and alternative hypotheses as $H_0 : \theta = 0$ and $H_1 : \theta > 0$, respectively. The null hypothesis H_0 states that the treatment has no effect, while the

alternative hypothesis H_1 states that the treatment has a significant effect. Statistical significance at a level of $\alpha \in (0,1)$ for the dataset $\mathcal{D}_S$ is defined as

$$p_{meta}(\mathcal{D}_S) = \Phi\left(-\sqrt{\sum_{i\in\mathcal{D}_S} w_i}\hat{\theta}\right) \leq \alpha, \tag{7}$$

where $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$, $\phi(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$, and $w_i = 1/\sigma_i^2$ or Equation (6) is used in the fixed- and random-effect models, respectively. The extension to two-sided tests is easy and is discussed later in Section 5.

2.1. Chance of Cherry-Picking in a Meta-Analysis

This section describes how the standard hypothesis testing procedure is no longer robust against selection bias due to the cherry-picking of studies using biased inclusion and/or exclusion criteria. We used similar techniques to those employed by Komiyama and Maehara [21] in the following derivation.

Theorem 1 guarantees that, under certain mild conditions, it is possible for meta-analysts to have sufficient statistical power to falsely conclude that a significant effect of the treatment of interest (Case 1) exists, even if the treatment has no actual effect. This is achieved by cherry-picking the subset $\mathcal{D}_S$ that provides the top-S smallest p-values.

Theorem 1. *For any $\alpha \in (0,1/2)$, $\delta \in (0,1)$, and $\epsilon \in (0,1/3)$, if $S/K \leq \epsilon$ and*

$$S \geq \max\left(\eta\left\{\frac{\Phi^{-1}(\alpha)}{\Phi^{-1}\left(\frac{1}{2}-\frac{\epsilon}{2}\right)}\right\}^2, \epsilon\left\{\frac{\log\left(\frac{1}{\delta}\right)}{2\left(\frac{1}{2}-\frac{3\epsilon}{2}\right)^2}-2\right\}\right),$$

with $\eta = \dfrac{w_{max}}{w_{min}}$, $w_{max} = \max_{i\in\mathcal{D}_S} w_i$ and $w_{min} = \min_{i\in\mathcal{D}_S} w_i$, then meta-analysts can select $\mathcal{D}_S$ such that $p_{meta}(\mathcal{D}_S) \leq \alpha$ with a probability of at least $1-\delta$.

Similarly, Theorem 2 guarantees that under certain conditions, it is also possible to falsely conclude that the treatment has an insignificant effect (Case 2), even if the treatment has an actual effect. This is achieved by cherry-picking the subset $\mathcal{D}_S$ that provides the top-S largest p-values.

Theorem 2. *For any $\alpha \in (0,1/2)$, $\delta \in (0,1)$, and $\epsilon \in (0,1)$, if $1-\epsilon \leq S/K \leq 1$ and*

$$\eta\left\{\frac{\Phi^{-1}(\alpha)}{\Phi^{-1}\left(1-\frac{\epsilon}{2}\right)}\right\}^2 \leq S < (1-\epsilon)\left\{\frac{\log\left(\frac{1}{\delta}\right)}{2(1-\epsilon/2)^2}-2\right\},$$

the meta-analysts can select $\mathcal{D}_S$ such that $p_{meta}(\mathcal{D}_S) \geq \alpha$ with a probability of at least δ.

Together, these theorems imply that meta-analysts have a chance to change the results of meta-analysis, regardless of real treatment effects, by cherry-picking an appropriate value for S. When S satisfies the conditions outlined in the theorems, readers or inspectors of the meta-analysis results can claim that the possibility of cherry-picking exists. In addition, now, we have assumed that meta-analysts cherry-pick the subset of studies $\mathcal{D}_S$ yielding the "top-S" (largest/smallest) p-values, which sometimes seems an unrealistic assumption because the actual meta-analysts might try to cherry-pick the subset of studies in a more arbitrary manner. However, it is noteworthy that even if meta-analysts cherry-pick an arbitrary subset of all studies such as the subset of studies with moderate p-values, these theorems are still valid because the current assumption of $\mathcal{D}_S$ is the most aggressive and worst setting, i.e., we assume that $\mathcal{D}_S$ provides the minimum/maximum p-value in the proof. Thus, the theorem still holds

even under the more relaxed assumption of cherry-picking moderate p-values. The proofs for the theorems can be found in Appendices A–C.

2.2. Extension to a Random-Effect Model

In the above section, we tentatively assumed that w_i was known, which corresponds to a fixed-effect model in a meta-analysis. However, we can also consider cases in which τ^2 in Equation (6) is estimated. In other words, we can estimate the between-study variance using the random-effect model. In practice, τ^2 is estimated from the data, frequently by using the method proposed by [3]. Given $\mathcal{D}_S$, the DerSimonian–Laird estimate of τ^2 is defined as

$$\max\left[0, \frac{\sum_{i\in\mathcal{D}_S} w_i(y_i - y_0)^2 - S + 1}{\sum_{i\in\mathcal{D}_S} w_i - \sum_{i\in\mathcal{D}_S} w_i^2 / \sum_{i\in\mathcal{D}_S} w_i}\right], \tag{8}$$

where $y_0 = \sum_{i\in\mathcal{D}_S} w_{i0}y_i / \sum_{i\in\mathcal{D}_S} w_{i0}$ and $w_{i0} = 1/\sigma_i^2$. Theorems 1 and 2 are nontrivial because this estimate depends on the choice of $\mathcal{D}_S$, and the selection of the top-S largest test statistics of $\sqrt{w_i}y_i$ depends on the estimate. These factors eliminate the simplicity of Theorems 1 and 2 and require a more sophisticated analysis. One possible approach is that, instead of using Equation (8), we replace $\mathcal{D}_S$ and S in Equation (8) with $\mathcal{D}_K$ and K, respectively. This corresponds to the situation where once τ^2 is estimated, it is regarded as a fixed constant in the model and the same discussion is applied with Theorems 1 and 2. The results of the random-effect models are examined in the simulation and application sections. In addition, in our future work, we plan to extend our results to cover cases in which τ^2 depends on the choice of $\mathcal{D}_S$.

3. Simulation Experiments

3.1. Simulation Settings

In this section, we describe Monte Carlo simulations that were implemented to demonstrate how sensitive the standard hypothesis test for meta-analyses is to the cherry-picking of studies, allowing meta-analysts to derive biased conclusions.

We considered both Case 1, where meta-analysts try to overstate the effectiveness of a treatment, despite there being no actual effect and Case 2, where meta-analysts try to understate the effectiveness of a treatment, despite there being an actual effect. The tunable parameters for the simulation scenarios were the number of cherry-picked studies $S = 2, \ldots, 30$, the proportion of cherry-picked studies among all studies $S/K \in \{1/3, 1/5, 1/10\}$, the true treatment effect $\theta \in \{0, 0.5, 1.0\}$, and the between-study variance $\tau^2 \in \{0, 0.01, 0.10, 0.50, 0.70\}$, where $\tau^2 = 0$ corresponds to the fixed-effect model and $\tau^2 > 0$ corresponds to the random-effect model. Additionally, $\theta = 0$ corresponds to Case 1 and $\theta > 0$ corresponds to Case 2. Following the approach described by Brockwell et al., (2001) [22], we simulated K independent studies. Each study has y_i and σ_i^2, where y_i and σ_i^2 are assumed to follow

$$y_i|\sigma_i \sim N(\theta, \sigma_i^2 + \tau^2), \;\; \sigma_i^2 \sim 0.25\chi_1^2.$$

The variances σ_i^2 were assumed to follow a χ_1^2 distribution, multiplied by 0.25 and truncated to an interval of $(0.009, 0.600)$, resulting in a mean within-study variance estimate of 0.17 [22]. Because τ^2 was varied from 0 to 0.70, the heterogeneity measure I^2 moved from 0% (no heterogeneity) to 80% (considerable heterogeneity). Throughout our simulations, we used $\alpha = 0.05$ as the type 1 error rate. Using these settings, we performed 1000 Monte Carlo simulations. In addition, values of $S > 30$, were examined, but they did not yield any notably different results. Therefore, we excluded the results for these settings.

3.2. Simulation Results

The simulation results revealed that when meta-analysts try to cherry-pick studies to change (or manipulate) pooled estimates and obtain their preferred conclusions, they have a chance of making it work in practice. Figures 1 and 2 present the simulation results

for Cases 1 and 2, respectively. They present the proportions of false conclusions (i.e., the proportion of 1000 iterations that succeeded in "flipping" the conclusion from significant to nonsignificant).

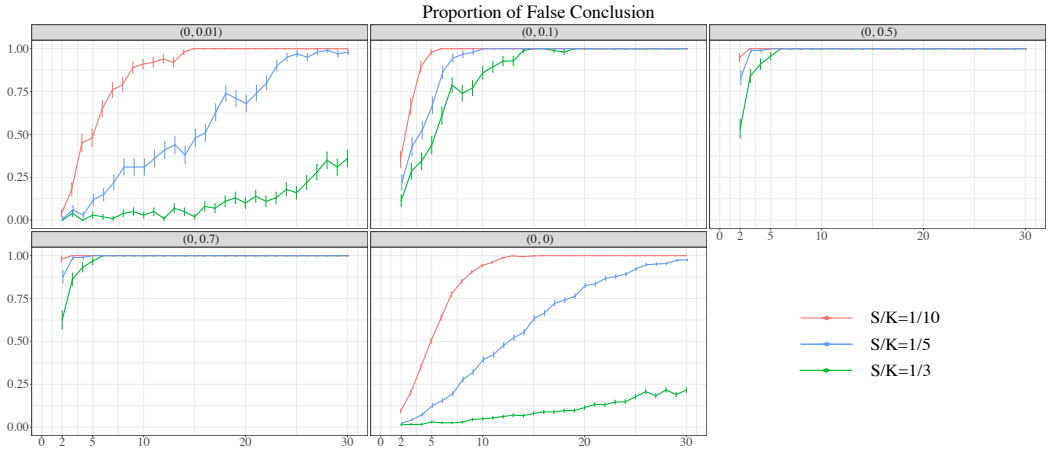

Figure 1. Simulation results for the standard hypotheses for Case 1, where meta-analysts overstate the effect of the treatment, regardless of there being no actual effect: Proportion of False Conclusions (i.e., type 1 error). (a, b) indicates $\theta = a$ and $\tau = b$.

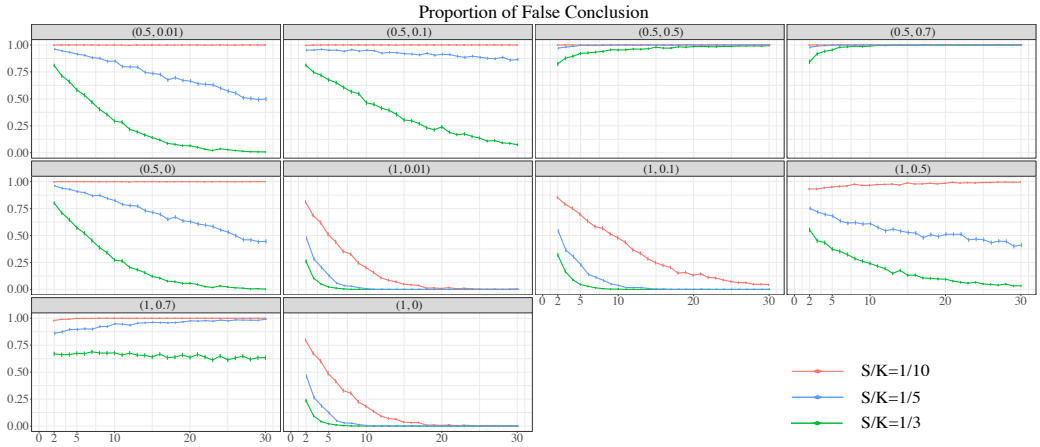

Figure 2. Simulation results for the standard hypotheses for Case 2, where meta-analysts understate the effect of the treatment, regardless of the actual effect: Proportion of False Conclusions (i.e., type 2 error). (a, b) indicates $\theta = a$ and $\tau = b$.

Figure 1 presents the results for Case 1, where the true treatment effect is $\theta = 0$. This indicates that, when using the random-effect model ($\tau^2 = 0.1, 0.5$), the proportion of false conclusions increases as S increases or S/K decreases in the standard hypothesis testing framework. In particular, when τ^2 is large, it is possible for meta-analysts to almost always cherry-pick studies to falsely conclude a significant treatment effect, despite there being no actual effect. When using the fixed-effect model ($\tau^2 = 0$), a similar tendency was observed: the proportion of false conclusions increased as S increased or S/K decreased. Figure 2 also presents the results for Case 2 where the true treatment effects are $\theta = 0.5$ and 1.0, respectively. This shows that meta-analysts still have a chance of cherry-picking studies to falsely conclude treatment insignificance, despite there being an actual effect. However, it shows different trends from Case 1: the proportion of false conclusions sometimes decreases as S increases. Especially when τ is small ($\tau = 0$ or 0.01) and θ is large, the standard hypothesis testing works well as S increases.

4. Medical Application Studies

This section shows how we can cherry-pick studies from two medical datasets. We emphasize that the original and subsequent analyses in the referenced articles were not cherry-picked. However, since cherry-picking is not reported in practice by definition, it is impossible to obtain real cherry-picked examples. Therefore, we made artificially cherry-picked situations from these real-world datasets, which are described in the following subsections.

4.1. Case 1 Example: Clinical Trials on the Effectiveness of Magnesium for Reducing the Mortality of Acute Myocardial Infarction Patients

We considered the results of randomized clinical trials (RCTs) that tested the effectiveness of intravenous magnesium for reducing the mortality following acute myocardial infarction (AMI). Because magnesium has been shown to protect ischemic myocytes from calcium overload, it is of significant interest to examine how magnesium can affect the mortality of ischemic heart disease patients. Teo et al. [23] conducted a meta-analysis using a fixed-effect approach based on seven studies (studies 1–7 in Figure 3), suggesting that magnesium has a significant effect on reducing the mortality of AMI patients. However, the results of a large trial (ISIS-4) indicated contradicting results, namely, that magnesium has no significant effect on reducing mortality [24]. We considered 16 studies that reported their summary statistics and the estimated odds ratio. They were extracted from Eggar and Smith [25], including ISIS-4 and the seven studies in Teo et al. [23]. Figure 3 presents the synthesized results when using the fixed-effect model presented by Teo et al. [23]. It is shown that magnesium has no significant effect on reducing mortality when considering all 16 studies; we obtained an estimated odds ratio (OR) (95% CI) of 0.994 (0.937, 1.055) and a p-value of 0.579 based on the one-sided test defined in Equation (3). Therefore, when using the 16 studies, there is a chance for meta-analysts to cherry-pick studies to obtain a biased treatment effect of magnesium (corresponds to Case 1). For example, when using the same set of studies as those used in Teo et al. [23], it appears that magnesium has a significant effect on mortality reduction with an estimated OR (95% CI) is 2.224 (1.401, 3.531) and p-value < 0.001. Therefore, relying on the results of Teo et al. [23] to conclude the effectiveness of magnesium (hypothetically) corresponds to Case 1.

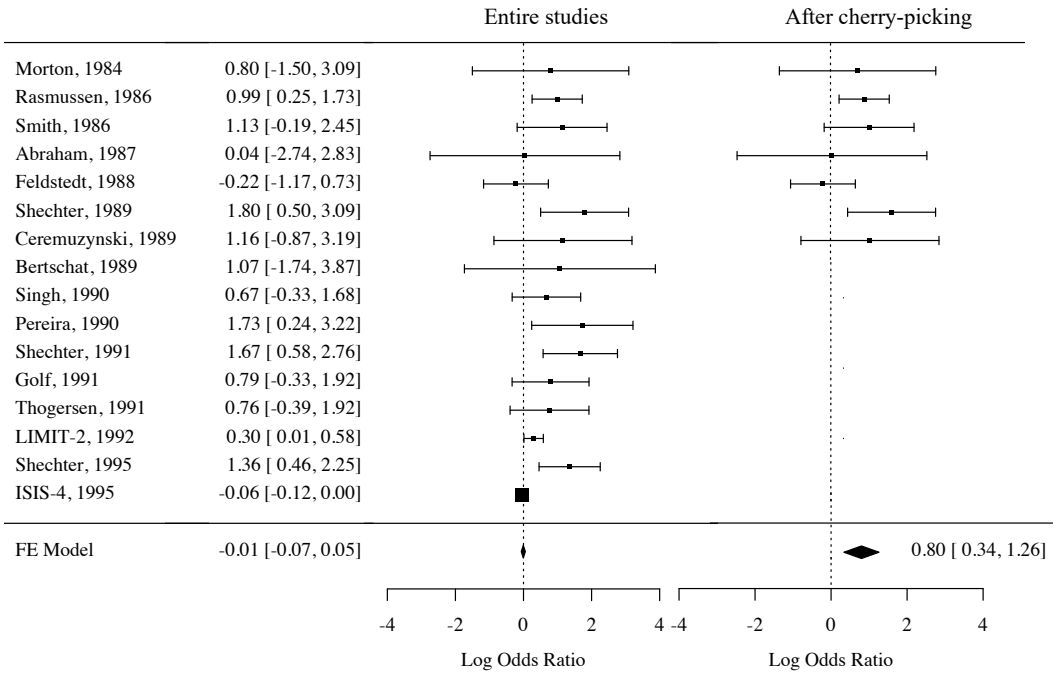

Figure 3. Meta-analysis of the results of 16 RCTs on the effectiveness of magnesium for reducing mortality following AMI.

4.2. Case 2 Example: Clinical Trials on the Effectiveness of St. John's Wort for Treating Depression

We considered the results of nine RCTs on the effectiveness of extracts of Hypericum perforatum (St. John's wort) for treating depression. Originally considered to be an effective treatment for depression, there have been mixed findings from several clinical trials comparing St. John's wort to a placebo. Linde et al. [26], from which we borrowed data, assessed a number of patients categorized as "responders" based on the Hamilton Rating Scale for Depression (HRSD). Notably, 17 studies were dropped from the group of studies used by Linde et al. [26] because they used a different version of the HRSD for assessing the degree of depression.

Figure 4 presents the results for the random-effect model. It simulates how meta-analysts can cherry-pick studies by selecting another definition of the treatment response (Def. 1: HRSD score reduction of at least 50% compared to baseline or HRSD score after therapy <10; Def. 2: HRSD reduction of at least 50% compared to baseline) to conclude the insignificant effectiveness of St. John's wort for depression, regardless of the actual effect (i.e., this case study corresponds to Case 2): St. John's wort provides a reduction in depression when considering all nine studies with an estimated OR (95% CI) of 1.467 (1.067, 2.016) and a p-value of 0.009. In contrast, when restricting our analysis to the subset of studies that applied only Def. 1 for the definition of a treatment response, we concluded that St. John's wort has a nonsignificant effect in reducing depression with an estimated OR (95% CI) of 1.458 (0.753, 2.822) and a p-value of 0.132.

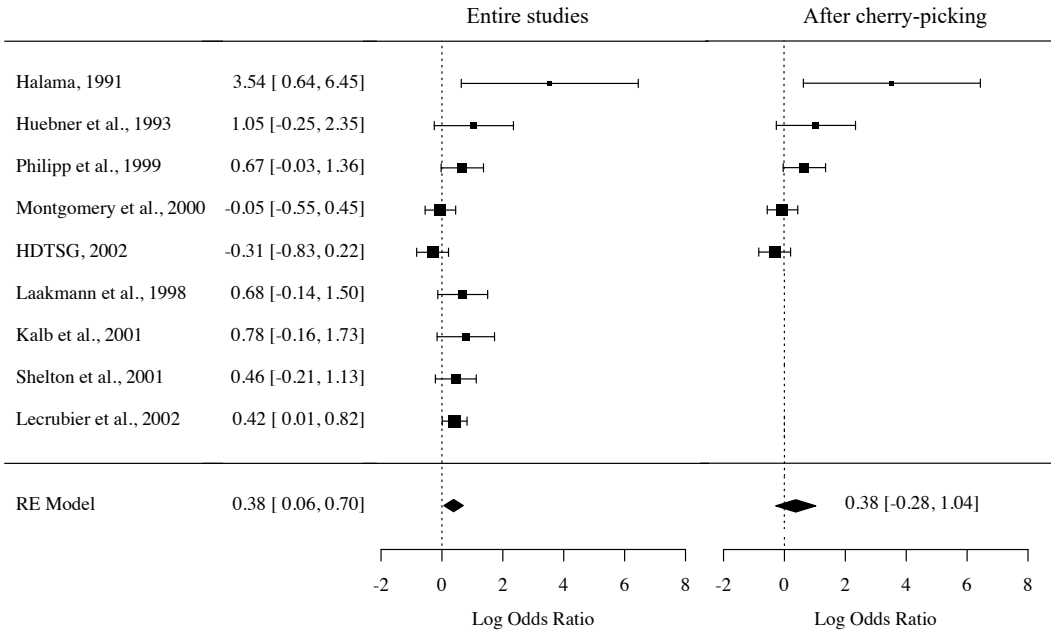

Figure 4. Meta-analysis of the results from nine RCTs on the effectiveness of St. John's wort for treating depression.

5. Discussion

The conclusions of any meta-analysis can be biased if meta-analysts intentionally or unintentionally cherry-pick a subset of all studies that lead to a desired favorable result. This is achieved by choosing beneficial inclusion and/or exclusion criteria. We theoretically assessed the conditions under which such cherry-picking is possible. To prevent cherry-picking in a meta-analysis, one solution is to mandate stricter adherence to Cochrane and other guidelines. This would require meta-analysts to register and publish their protocol before carrying out the primary meta-analysis. In addition, a more advanced mechanism would be necessary to verify the inclusion/exclusion criteria that were not initially included in the protocol but were subsequently added. The R code is provided in a GitHub repository (https://github.com/kingqwert/R/tree/master/metaCherry/, accessed on 2 March 2023)

and will be hosted on the *R* CRAN repository (https://www.r-project.org/, accessed on 2 March 2023) in the near future, allowing others to apply our method easily.

Extensive Monte Carlo simulations were conducted to illustrate that the standard meta-analysis method could be subject to cherry-picking, leading to biased results. The chance of cherry-picking is remarkably high, especially when S is small. Furthermore, two real data analysis problems were simulated to provide new insights into the results of RCTs on the effectiveness of magnesium on AMI and St. John's wort on depression. We demonstrated that it is easy to obtain favorable, i.e., biased, conclusions by cherry-picking studies based on biased inclusion and/or exclusion criteria. We encourage the re-evaluation of our approach using other datasets.

We demonstrated that meta-analysts can cherry-pick a subset of studies by modifying inclusion and/or exclusion criteria. However, this type of cherry-picking should not be taken too literally: the theorems presented in this study can be applied to any type of cherry-picking if information regarding K, S, and w_i is available. In addition, we analyzed the case of cherry-picking from a 'subset' of studies (i.e., the case of $S < K$). It is trivial to extend this analysis to the case of $K < S$, where meta-analysts use unsuitable inclusion and/or exclusion criteria to increase the total number of studies to obtain a favorable conclusion. Similarly, although we focused on the case of one-sided right-tailed hypothesis testing in this study, it is simple to extend our results to (i) the one-sided left-tailed hypothesis case ($H_0 = 0$ and $H_0 < 0$) by using $p_{meta} = \Phi(\hat{\theta}\sqrt{\sum_{i\in\mathcal{D}_S} w_i})$, and (ii) the two-sided hypothesis case ($H_0 = 0$ and $H_0 \neq 0$) by using $p_{meta} = 2\Phi(-\hat{\theta}\sqrt{\sum_{i\in\mathcal{D}_S} w_i})$ instead of Equation (7). In addition, the assumption of cherry-picking the top-S results is sometimes unrealistic, and actual meta-analysts might try to cherry-pick the subset of studies in a more arbitrary manner. However, we note again that, as discussed in Section 2.1, the theorems are still valid, even if meta-analysts cherry-pick an arbitrary subset of all studies.

Similar to most published studies on meta-analyses, the within-study variance σ_i^2 in Equation (1) was assumed to be known, ignoring the fact that it must be estimated in practice. If the estimated σ_i^2 and τ^2 values are used to define the p-value, it no longer follows a standard normal distribution under the null hypothesis [27–29], eliminating the simplicity of our theorems. In such cases, a more complicated asymptotic analysis would be required. Furthermore, there have been many previous attempts to formulate a "publication bias" using p-values [30,31]. It would be worthwhile to consider both selection and publication biases simultaneously by using the proposed framework for hypothesis testing and its associated p-value. However, further discussion about the conceptual difference between the publication bias and selection bias due to cherry-picking is required.

Author Contributions: Conceptualization, D.Y.; methodology, D.Y.; formal analysis, D.Y. and B.R.; data curation, D.Y.; writing—original draft preparation, D.Y. and B.R.; writing—review and editing, D.Y. and B.R.; visualization, D.Y.; supervision, D.Y.; project administration, D.Y.; funding acquisition, D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by JST, PRESTO Grant Number JPMJPR21RC, Japan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

RCT randomized clinical trials

Appendix A

Lemma A1. *Let $t > 0$ and $X \sim Beta(\alpha, \beta)$, where $Beta(\alpha, \beta)$ is a beta distribution with shape parameters α and β. Then,*

$$P\left[X - \frac{\alpha}{\alpha+\beta} > t\right] \leq \exp(-2(\alpha+\beta+1)t^2).$$

Proof. Marchal and Arbel [32] demonstrated that the beta distribution $Beta(\alpha, \beta)$ is σ^2-sub-Gaussian with a mean $\alpha/(\alpha+\beta)$ and $\sigma^2 = 1/(4(\alpha+\beta+1))$. It is well known that for a σ^2-sub-Gaussian random variable X with a mean μ and variance factor σ^2, $P[X - \mu > t] \leq \exp(-t^2/(2\sigma^2))$ holds for any $t > 0$. □

Appendix B. Proof of Theorem 1

Proof. We will follow the techniques presented in Komiyama and Maehara [21] for our proofs. An adversarial meta-analyst is assumed to cherry-pick studies that provide the top-S smallest test statistics of $-\sqrt{w_i}y_i$ (this is equivalent to the top-S smallest p-values). Assume $\Phi^{-1}(\alpha) = x$ such that $\Phi(x) = \alpha$ and $\{-\sqrt{w_i}y_i\}_{i\in\mathcal{D}_S}$ are sorted as $\sqrt{w_1}y_1 \geq \sqrt{w_2}y_2 \geq \cdots \geq \sqrt{w_S}y_S$, which is equivalent to $p_1 \leq p_2 \leq \cdots \leq p_S$.

From Equations (3) and (7), we have

$$\frac{w_i y_i}{\sum_{i\in\mathcal{D}_S} w_i} = -\frac{\sqrt{w_i}}{\sum_{i\in\mathcal{D}_S} w_i}\Phi^{-1}(p_i), \quad \hat{\theta} \geq -\frac{\Phi^{-1}(\alpha)}{\sqrt{\sum_{i\in\mathcal{D}_S} w_i}}. \tag{A1}$$

Note that, from Equations (2) and the assumption, we have

$$\hat{\theta} = \frac{\sum_{i\in\mathcal{D}_S} w_i y_i}{\sum_{i\in\mathcal{D}_S} w_i} = -\frac{\sum_{i\in\mathcal{D}_S} \sqrt{w_i}\Phi^{-1}(p_i)}{\sum_{i\in\mathcal{D}_S} w_i} \geq -\frac{\sum_{i\in\mathcal{D}_S} \sqrt{w_i}}{\sum_{i\in\mathcal{D}_S} w_i}\Phi^{-1}(p_S). \tag{A2}$$

Therefore, for Equations (A1) and (A2), the sufficient condition can be written as

$$-\frac{\sum_{i\in\mathcal{D}_S} \sqrt{w_i}}{\sum_{i\in\mathcal{D}_S} w_i}\Phi^{-1}(p_S) \geq -\frac{\Phi^{-1}(\alpha)}{\sqrt{\sum_{i\in\mathcal{D}_S} w_i}}, \tag{A3}$$

which is equivalent to

$$p_S \leq \Phi\left(\frac{\sqrt{\sum_{i\in\mathcal{D}_S} w_i}}{\sum_{i\in\mathcal{D}_S} \sqrt{w_i}}\Phi^{-1}(\alpha)\right). \tag{A4}$$

Additionally, we have

$$\begin{aligned}\Phi\left(\frac{\sqrt{\sum_{i\in\mathcal{D}_S} w_i}}{\sum_{i\in\mathcal{D}_S} \sqrt{w_i}}\Phi^{-1}(\alpha)\right) &\geq \Phi\left(\frac{\sqrt{S w_{max}}}{S\sqrt{w_{min}}}\Phi^{-1}(\alpha)\right) \quad (\Phi^{-1}(\alpha) < 0 \text{ since } \alpha \in (0, 1/2)) \\ &\geq \frac{1}{2}(1-\epsilon), \quad \text{(the assumption)}\end{aligned}$$

where $w_{max} = \max_{i\in\mathcal{D}_S} w_i$ and $w_{min} = \min_{i\in\mathcal{D}_S} w_i$. Therefore, the sufficient condition can be written as

$$p_S \leq \frac{1}{2}(1-\epsilon). \tag{A5}$$

In contrast, it is assumed that each p_i $(i = 1, \ldots, K)$ is an i.i.d random sample from the uniform distribution $U(0,1)$. Then, because meta-analysts are supposed to pick the top-S smallest p_i $(i = 1, \ldots, S)$ values and $\{p_i\}_{i=1}^{S}$ is ordered as $p_1 \leq p_2 \leq \cdots \leq p_S$, we have

$$p_S \sim Beta(S, K - S + 1),$$

where $Beta(\alpha, \beta)$ is a beta distribution with the shape parameters α and β. Therefore, based on Lemma 1 and the assumption above, the following statement holds with a probability of $1 - \delta$:

$$p_S \leq \frac{S}{K+1} + \sqrt{\frac{\log(1/\delta)}{2(K+2)}} < \frac{S}{K} + \sqrt{\frac{\log(1/\delta)}{2(S/\epsilon+2)}} \leq \frac{1}{2}(1-\epsilon). \tag{A6}$$

By rearranging Equations (A5) and (A6), we obtain the claim of the theorem. □

Appendix C. Proof of Theorem 2

Proof. Because the two cases discussed in Theorems 1 and 2 are somewhat complementary, this proof is similar to the proof of Theorem 1 and follows similar reasoning. An adversarial meta-analyst is assumed to cherry-pick studies that provide the top-S largest test statistics of $-\sqrt{w_i}y_i$ (which is equivalent to the top-S largest p-values). It is assumed that $\Phi^{-1}(\alpha) = x$, such that $\Phi(x) = \alpha$ and $\{-\sqrt{w_i}y_i\}_{i \in \mathcal{D}_S}$ are sorted as $\sqrt{w_1}y_1 \leq \sqrt{w_2}y_2 \leq \cdots \leq \sqrt{w_S}y_S$, which is equivalent to $p_1 \geq p_2 \geq \cdots \geq p_S$. Now, we wish to show that

$$p_{meta}(\mathcal{D}_S) = \Phi\left(-\sqrt{\sum_{i \in \mathcal{D}_S} w_i}\hat{\theta}\right) \geq \alpha. \tag{A7}$$

From Equations (3) and (7), we have

$$\frac{w_i y_i}{\sum_{i \in \mathcal{D}_S} w_i} = -\frac{\sqrt{w_i}}{\sum_{i \in \mathcal{D}_S} w_i}\Phi^{-1}(p_i), \quad \hat{\theta} \leq -\frac{\Phi^{-1}(\alpha)}{\sqrt{\sum_{i \in \mathcal{D}_S} w_i}}. \tag{A8}$$

Note that from Equation (2) and the assumption, we have

$$\hat{\theta} = \frac{\sum_{i \in \mathcal{D}_S} w_i y_i}{\sum_{i \in \mathcal{D}_S} w_i} = -\frac{\sum_{i \in \mathcal{D}_S} \sqrt{w_i}\Phi^{-1}(p_i)}{\sum_{i \in \mathcal{D}_S} w_i} \leq -\frac{\sum_{i \in \mathcal{D}_S} \sqrt{w_i}}{\sum_{i \in \mathcal{D}_S} w_i}\Phi^{-1}(p_S). \tag{A9}$$

Therefore, for Equations (A8) and (A9), the sufficient condition can be written as

$$-\frac{\sum_{i \in \mathcal{D}_S} \sqrt{w_i}}{\sum_{i \in \mathcal{D}_S} w_i}\Phi^{-1}(p_S) \leq -\frac{\Phi^{-1}(\alpha)}{\sqrt{\sum_{i \in \mathcal{D}_S} w_i}}, \tag{A10}$$

which is equivalent to

$$p_S \geq \Phi\left(\frac{\sqrt{\sum_{i \in \mathcal{D}_S} w_i}}{\sum_{i \in \mathcal{D}_S} \sqrt{w_i}}\Phi^{-1}(\alpha)\right). \tag{A11}$$

Additionally, we have

$$\begin{aligned}\Phi\left(\frac{\sqrt{\sum_{i\in\mathcal{D}_S} w_i}}{\sum_{i\in\mathcal{D}_S}\sqrt{w_i}}\Phi^{-1}(\alpha)\right) &\leq \Phi\left(\frac{\sqrt{Sw_{min}}}{S\sqrt{w_{max}}}\Phi^{-1}(\alpha)\right) \quad (\Phi^{-1}(\alpha)<0 \text{ since } \alpha\in(0,1/2))\\ &= \Phi\left(\sqrt{\frac{w_{min}}{Sw_{max}}}\Phi^{-1}(\alpha)\right)\\ &\leq 1-\frac{\epsilon}{2}, \qquad \text{(the assumption)}\end{aligned}$$

where $w_{max} = \max_{i\in\mathcal{D}_S} w_i$ and $w_{min} = \min_{i\in\mathcal{D}_S} w_i$. Therefore, the sufficient condition for (A11) can be written as

$$p_S \geq 1-\frac{\epsilon}{2}. \tag{A12}$$

By contrast, it is assumed that each p_i $(i=1,\ldots,K)$ is an i.i.d random sample from the uniform distribution $U(0,1)$. Then, as meta-analysts are supposed to pick the top-S largest p_i $(i=1,\ldots,S)$ and $\{p_i\}_{i=1}^{S}$ is ordered as $p_1 \geq p_2 \geq \cdots \geq p_S$, the following statement holds:

$$p_S \sim Beta(K-S+1,S).$$

Therefore, based on Lemma 1 and the assumption above, the following statement holds with a probability of δ:

$$\begin{aligned}p_S &\geq \frac{K-S+1}{K+1}+\sqrt{\frac{\log(1/\delta)}{2(K+2)}}\\ &> \sqrt{\frac{\log(1/\delta)}{2(K+2)}}\\ &> \sqrt{\frac{\log(1/\delta)}{2\left(\frac{S}{1-\epsilon}+2\right)}}\\ &\geq 1-\frac{\epsilon}{2}.\end{aligned} \tag{A13}$$

By rearranging Equations (A12) and (A13), we obtain the claim of the theorem. □

References

1. Higgins, J.P.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M.J.; Welch, V.A. *Cochrane Handbook for Systematic Reviews of Interventions*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
2. Whitehead, A. *Meta-Analysis of Controlled Clinical Trials*; John Wiley & Sons: Hoboken, NJ, USA, 2002; Volume 7.
3. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **1986**, *7*, 177–188. [CrossRef] [PubMed]
4. Egger, M.; Smith, G.D.; Schneider, M.; Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **1997**, *315*, 629–634. [CrossRef] [PubMed]
5. Hopewell, S.; Clarke, M.J.; Lefebvre, C.; Scherer, R.W. Handsearching versus electronic searching to identify reports of randomized trials. In *Cochrane Database of Systematic Reviews*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
6. Ioannidis, J.P. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* **1998**, *279*, 281–286. [CrossRef]
7. Higgins, J.P.; Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **2002**, *21*, 1539–1558. [CrossRef] [PubMed]
8. Yoneoka, D.; Henmi, M. Meta-analytical synthesis of regression coefficients under different categorization scheme of continuous covariates. *Stat. Med.* **2017**, *36*, 4336–4352. [CrossRef]
9. Yoneoka, D.; Henmi, M.; Sawada, N.; Inoue, M. Synthesis of clinical prediction models under different sets of covariates with one individual patient data. *BMC Med. Res. Methodol.* **2015**, *15*, 1. [CrossRef] [PubMed]
10. Sutton, A.J.; Abrams, K.R.; Jones, D.R.; Jones, D.R.; Sheldon, T.A.; Song, F. *Methods for Meta-Analysis in Medical Research*; Wiley: Hoboken, NJ, USA, 2000.

11. Hedges, L.V.; Olkin, I. *Statistical Methods for Meta-Analysis*; Academic Press: Cambridge, MA, USA, 2014.
12. Borenstein, M.; Hedges, L.V.; Higgins, J.; Rothstein, H.R. *Introduction to Meta-Analysis*; John Wiley & Sons: New York, NY, USA, 2009.
13. Cooper, H.; Hedges, L.V.; Valentine, J.C. *The Handbook of Research Synthesis and Meta-Analysis*; Russell Sage Foundation: New York, NY, USA, 2009.
14. Page, M.J.; Shamseer, L.; Altman, D.G.; Tetzlaff, J.; Sampson, M.; Tricco, A.C.; Catala-Lopez, F.; Li, L.; Reid, E.K.; Sarkis-Onofre, R.; et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: A cross-sectional study. *PLoS Med.* **2016**, *13*, e1002028. [CrossRef] [PubMed]
15. Tawfik, G.M.; Giang, H.T.N.; Ghozy, S.; Altibi, A.M.; Kandil, H.; Le, H.H.; Eid, P.S.; Radwan, I.; Makram, O.M.; Hien, T.T.T.; et al. Protocol registration issues of systematic review and meta-analysis studies: A survey of global researchers. *BMC Med. Res. Methodol.* **2020**, *20*, 213. [CrossRef] [PubMed]
16. Clarke, M.J.; Stewart, L.A. Systematic Reviews: Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? *BMJ* **1994**, *309*, 1007–1010. [CrossRef]
17. Ahmed, I.; Sutton, A.J.; Riley, R.D. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: A database survey. *BMJ* **2012**, *344*, d7762. [CrossRef] [PubMed]
18. Grégoire, G.; Derderian, F.; Le Lorier, J. Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias? *J. Clin. Epidemiol.* **1995**, *48*, 159–163. [CrossRef]
19. Egger, M.; Smith, G.D. Meta-analysis bias in location and selection of studies. *BMJ* **1998**, *316*, 61–66. [CrossRef] [PubMed]
20. Knight, J. Null and void. *Nature* **2003**, *422*, 554–555. [CrossRef] [PubMed]
21. Komiyama, J.; Maehara, T. A Simple Way to Deal with Cherry-picking. *arXiv* **2018**, arXiv:1810.04996.
22. Brockwell, S.E.; Gordon, I.R. A comparison of statistical methods for meta-analysis. *Stat. Med.* **2001**, *20*, 825–840. [CrossRef]
23. Teo, K.K.; Yusuf, S.; Collins, R.; Held, P.H.; Peto, R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ* **1991**, *303*, 1499–1503. [CrossRef] [PubMed]
24. International Study Group of Infarct Survival. ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. *Lancet* **1995**, *345*, 669–85. [CrossRef]
25. Eggar, M.; Smith, G.D. Misleading meta-analysis. Lessons from "an effective, safe, simple" intervention that wasn't. *BMJ* **1995**, *310*, 752–754.
26. Linde, K.; Berner, M.; Egger, M.; Mulrow, C. St John's wort for depression: Meta-analysis of randomised controlled trials. *Br. J. Psychiatry* **2005**, *186*, 99–107. [CrossRef]
27. Malzahn, U.; Böhning, D.; Holling, H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika* **2000**, *87*, 619–632. [CrossRef]
28. Cornell, J.E.; Mulrow, C.D.; Localio, R.; Stack, C.B.; Meibohm, A.R.; Guallar, E.; Goodman, S.N. Random-effects meta-analysis of inconsistent effects: A time for change. *Ann. Intern. Med.* **2014**, *160*, 267–270. [CrossRef] [PubMed]
29. Langan, D.; Higgins, J.P.; Jackson, D.; Bowden, J.; Veroniki, A.A.; Kontopantelis, E.; Viechtbauer, W.; Simmonds, M. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res. Synth. Methods* **2019**, *10*, 83–98. [CrossRef] [PubMed]
30. Henmi, M.; Copas, J.B.; Eguchi, S. Confidence intervals and P-values for meta-analysis with publication bias. *Biometrics* **2007**, *63*, 475–482. [CrossRef] [PubMed]
31. Copas, J.; Jackson, D. A bound for publication bias based on the fraction of unpublished studies. *Biometrics* **2004**, *60*, 146–153. [CrossRef] [PubMed]
32. Marchal, O.; Arbel, J On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electron. Commun. Probab.* **2017**, *22* 1–14. [CrossRef]