

This article has been accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination, and proofreading processes, that may lead to differences between this version and the version of record.

Please cite this article as <https://doi.org/10.4097/kja.26020>

Epub ahead of print

Correcting what cannot be corrected: rethinking publication bias analysis methods in clinical meta-analyses

Hyun Kang, M.D., Ph.D., M.P.H.

Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine,
Seoul, Republic of Korea

Corresponding author:

Hyun Kang, M.D., Ph.D., M.P.H.

Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine, 84
Heukseok-ro, Dongjak-Gu, Seoul 06911, Republic of Korea

E-mail: roman00@naver.com; Tel:82-10-8761-4426

Running title: Publication bias analysis methods

Previous presentations: Not applicable

Conflict of interest: Hyun Kang has been a Statistical Rounds board member of the *KJA* since 2013. However, he was not involved in any process of review for this article, including peer reviewer selection, evaluation, or decision-making. No other potential conflict of interest relevant to this article was reported.

Funding: Not applicable.

Acknowledgments: Not applicable.

IRB number: Not applicable.

Research registration number: Not applicable.

Abstract

Methods to assess and adjust for publication bias are often presented as tools to correct distorted evidence in meta-analyses. However, statistical adjustment cannot recover information that was selectively generated, reported, or disseminated. Clinical evidence syntheses frequently rely on small or selective sets of trials and are characterized by substantial heterogeneity, multiple outcomes and time points, and complex dissemination pathways. Publication bias analysis methods are thus prone to over-interpretation and may yield conflicting conclusions. Therefore, they should be understood as an inferential process that links detection, model-based adjustment, and interpretation under explicit and unverifiable assumptions. We review classical methods to detect publication bias, including funnel plots, tests of small-study effects, and P-value-based approaches, and demonstrate their essential role as stress tests of model adequacy rather than as definitive detectors of publication bias. We then examine widely used methods to adjust for publication bias, such as trim-and-fill, selection models, regression-based approaches relating the effect size to study precision, and the Bayesian approach, clarifying their key assumptions and typical failure modes. Using a worked example, we illustrate how applying different publication bias adjustment methods to the same evidence base can yield divergent adjusted effects, emphasizing their assumption dependence. We additionally identify common misuses, propose a framework for evaluations, and discuss emerging challenges related to preprints, umbrella reviews, and AI-assisted evidence synthesis. This review thus aims to help align the strength of clinical conclusions with the robustness or fragility of the underlying data, with direct implications for authors, reviewers, and editors.

Keywords: Biostatistics, Evidence-based practice, Evidence Synthesis [Publication type], Meta-analysis as topic, Publication bias, Systematic review [Publication type].

1 **Introduction**

2 Publication bias is widely recognized as a major threat to the validity of meta-analyses in
3 clinical medicine [1-3]. It is often described as a tendency to publish only statistically significant or
4 favorable results; however, in practice, it reflects a much broader set of selective processes across
5 the entire research ecosystem. Selectivity can arise during topic prioritization, protocol design,
6 choice of outcomes and time points, analytic decisions, reporting, journal submission, peer review,
7 and citation. By the time evidence is synthesized, the available studies may thus represent a
8 systematically filtered subset of all research that was conducted [2, 4].

9 Meta-analyses in clinical medicine are particularly vulnerable to such distortions. Many
10 clinical questions are addressed by a small number of trials with variable outcome definitions,
11 inconsistent follow-up windows, and heterogeneous patient populations and interventions. Moreover,
12 research results now appear not only in journals but also in trial registries, preprints, conference
13 abstracts, press releases, and regulatory documents, each with their own selection pressures and
14 timelines [5, 6]. Under these conditions, making the assumption that studies are missing solely
15 because their results are nonsignificant is rarely realistic. Standard methods for detecting publication
16 bias such as funnel plots and small-study effect tests are often applied in situations that differ
17 substantially from the ideal settings for which they were originally developed.

18 Recent methodological work shows that publication bias can distort not only the apparent
19 magnitude of pooled effects, but also the consistency and trustworthiness of the evidence,
20 particularly when data are sparse and heterogeneity is substantial [7-11]. At the same time, the
21 growing list of methods available to assess and adjust for publication bias, from trim-and-fill and
22 regression-based corrections to selection models, P-value-based approaches, and the Bayesian
23 publication bias model, has made it more difficult for analysts to interpret conflicting signals and

multiple “corrected” estimates [7, 12-15]. The proliferation of these methods has sharpened the central tension that analyses of publication bias are often treated as automatic correction tools that recover a hidden “true effect,” even though all available methods to adjust for publication bias depend on strong, untestable assumptions about the selection process [8, 9, 16, 17].

In this review, publication bias analysis is therefore reframed as a single, coherent inferential process that integrates the methods used to detect and adjust for publication bias and their interpretations rather than as a checklist of disconnected techniques or a means to obtain one definitive bias-corrected effect [8, 16, 17]. The focus is on how each method encodes specific assumptions about which studies are more likely to appear in the literature, how these assumptions interact with between-study heterogeneity and selective outcome reporting, and the extent that clinical conclusions depend on these assumptions [7, 14, 15]. Drawing on empirical and methodological work on sensitivity analysis, selection models, and model-averaging, the methods used to detect and adjust for publication bias and interpretive guidance are combined into a single framework in this review, with practical recommendations for authors, reviewers, editors, and umbrella review teams [7, 14, 15]. Rather than aiming to recover a single “true” effect size, this perspective treats methods to assess and adjust for publication bias as tools for disciplined inference that help to align the strength and wording of clinical conclusions with the robustness and stability of the underlying evidence [4-6, 9]. In this sense, publication bias methods are best viewed not as mechanical correction tools, but as structured sensitivity analyses that clarify how clinical conclusions may depend on assumptions about missing evidence.

Publication bias in the research ecosystem

Viewing publication bias only as the end result of selectively publishing significant studies

obscures the larger system that produces it. Modern research environments reward novelty, clear positive messages, and statistically significant findings, so researchers are encouraged to choose questions that are likely to “work,” select outcomes and time points that make effects easier to detect, and emphasize favorable analyses when reporting results. Journals, funders, and media outlets pay more attention to studies that appear impactful, and citation patterns amplify these studies further, making them much more visible to later reviewers and meta-analysts [4-6, 18].

“Missingness” is rarely a simple yes/no issue. Not only are some entire studies not published, but bias often arises from selective reporting within published studies, for example, presenting primary but not secondary outcomes, highlighting favorable subgroups, choosing particular analytic models, or focusing on follow-up times that yield statistically significant or directionally desirable results [8, 19]. However, many standard methods to detect publication bias were developed under the assumption that entire studies are missing, rather than that certain outcomes or analyses are selectively reported, so they may respond only weakly or in misleading ways when the main problems are outcome switching, multiplicity, or data-driven analysis choices [8, 19]. Consequently, meta-analyses in clinical medicine can inherit an evidence base that is distorted both by the studies that are available and by the information reported in those studies, and no statistical adjustment at the synthesis stage can fully reconstruct information that was never transparently generated, recorded, or reported [6, 8, 9, 17, 18].

Publication bias is therefore best understood as an ecosystem-level phenomenon driven by the incentives and routines that shape the questions that are asked, how studies are designed and analyzed, the framing of results, and how they circulate through scientific and clinical communities [5, 6, 8, 16]. Within this ecosystem, methods designed to detect and adjust for publication bias are better viewed as tools for testing the robustness of findings under plausible selection and reporting mechanisms, rather than as algorithms that can recover a single underlying truth [8, 9, 16]. This

ecosystem perspective naturally motivates a distinction between detection, adjustment, and interpretation, and underpins the integrated inferential framework developed in the following sections [8, 9, 16].

Beyond classic publication bias, a wide range of dissemination- and reporting-related mechanisms can distort the body of evidence available for meta-analyses. These include the delayed or selective publication of entire studies, selective reporting of outcomes and analyses within studies, preferential citation of positive results, and differential visibility of findings across journals, preprints, registries, conference abstracts, and regulatory documents [20, 21]. Because these mechanisms can generate similar statistical patterns, such as small-study effects or funnel plot asymmetry, while calling for different interpretive responses, distinguishing them conceptually before introducing specific methods to detect and adjust for publication bias is a helpful strategy. **Table 1** summarizes a working taxonomy of dissemination-related and reporting biases that are most relevant to meta-analyses in clinical medicine, extending the notion of “publication bias” to a broader ecosystem of selective dissemination.

A well-known example is the Vioxx (rofecoxib) scandal [22-24]. Rofecoxib, a selective COX-2 inhibitor widely used for arthritis and pain, was withdrawn from the market after randomized trials and observational studies revealed an increased risk of myocardial infarction and stroke [22-24]. Subsequent investigations of regulatory submissions and company documents showed that some cardiovascular events had been incompletely reported, analyzed in non-prespecified ways, or published with substantial delay, and that unfavorable findings were less visible than positive efficacy results [22-24]. Conceptually, this episode maps directly onto the taxonomy shown in **Table 1**, combining publication and time-lag bias, outcome-reporting bias, selective analysis, and location/citation bias into a single ecosystem-level distortion of the evidence base.

1 Statistical detection of publication bias

2 *Funnel plots as stress tests, not detectors*

3 Funnel plots are often used as the initial check for publication bias in meta-analyses in clinical
4 medicine because they are simple to make and easy to read [12, 16, 25, 26]. They plot study effect
5 estimates against a measure of precision and assume an ideal situation: if studies are an unbiased
6 sample and true effects are broadly similar, less precise studies should be spread roughly
7 symmetrically around the pooled effect. **Fig. 1A** illustrates this approach using the tropisetron dataset,
8 with each circle representing an individual study positioned according to its log risk ratio (x-axis) and
9 standard error (y-axis), with more precise studies toward the bottom. The contour lines demarcate
10 regions of statistical significance ($P < 0.05$ and $P < 0.01$), providing additional context for interpreting
11 asymmetry. Visual inspection reveals an apparent deficit of small, unfavorable studies in the lower-
12 right quadrant and a cluster of small studies reporting large protective effects on the left side, creating
13 an asymmetric funnel shape that could initially suggest publication bias.

14 However, when the same data are examined with additional structures such as stratification
15 by methodological quality (**Fig. 1B**), the apparent asymmetry becomes less consistent with selective
16 non-publication. In this example, several small, low-quality trials are clustered on the left side of the
17 funnel with relatively large protective effects, whereas larger trials with higher methodological quality
18 tend to lie closer to the null. Such a pattern can arise even when no studies are missing: studies with
19 a higher risk of bias are more likely to overestimate treatment effects, particularly when sample sizes
20 are small and outcome assessment is less rigorous. In this setting, the funnel plot may appear
21 asymmetric despite being complete, reflecting an exaggeration of the risk-of-bias-related effect rather
22 than publication bias.

23 In practice, true effects can differ according to baseline risk, disease severity, co-interventions,

adherence, or outcome measurement methods, which are factors that often correlate with study size or design. Therefore, funnel plot asymmetry can be caused by real effect modifications, systematic differences in methods, or differences in outcome definitions even when no studies are missing [2, 18]. Funnel plots are also difficult to interpret when only a few studies are available [27-29]. Random variation or a single influential trial can create patterns that resemble genuine asymmetry [27-29]. Conversely, a funnel plot may appear roughly symmetric even in the presence of substantial selective reporting, especially when the main source of bias is within-study selection (e.g., outcome switching or selective choice of follow-up time) rather than non-publication of entire studies [8, 19].

For these reasons, funnel plots are best viewed as stress tests of how well a simple model with homogeneous effects and no selection fits the data and not as tools that identify a specific mechanism. They indicate that something is wrong with the simple model and that further investigation is needed; however, they cannot clarify whether publication bias, heterogeneity, or other factors are responsible [6, 18].

Egger's and Begg's tests: small-study effects are not publication bias

Regression- and rank-based tests, such as Egger's regression and Begg's rank-correlation tests, turn funnel plot asymmetry into a statistical test by examining whether effect sizes are related to study precision [30, 31]. These tests are often interpreted as direct methods for detecting publication bias. However, in reality, they detect small-study effects, a broader pattern in which smaller studies tend to report different (often larger) effects than larger studies.

Small-study effects do not indicate a single cause; publication bias is only one possible explanation. Other common reasons include real differences in effects across study settings or patient populations; systematic differences in study designs or methodological quality; selective reporting of outcomes; exploratory analyses or the early termination of small trials; and timing effects, in which

1 small studies are conducted early, followed by later, larger, and more definitive trials.

2 Because of these alternative explanations, a statistically significant Egger's or Begg's test
3 should be interpreted as evidence of an unexplained relationship between effect size and study
4 precision and not as proof that publication bias is present. These tests are also sensitive to data
5 limitations: when heterogeneity is high, false-positive results may be produced, and when only a few
6 studies are available, they may lack power and fail to detect real problems. In some cases, the result
7 can be driven largely by a single influential study [2, 18].

8 Conversely, a non-significant asymmetry test does not rule out publication bias. This is
9 particularly true when the number of studies is small, heterogeneity is substantial, or selective
10 reporting occurs within studies rather than through non-publication of whole trials [7, 8, 19]. To use
11 Egger's and Begg's tests appropriately, analysts should consider alternative explanations for small
12 study effects and interpret test results alongside other information such as risk-of-bias assessments,
13 study design features, and evidence from trial registries [2, 9, 18].

14 ***P-value-based methods: p-curve, p-uniform, and evidential value under selective reporting***

15 P-value-based methods assess publication bias by examining the distribution of statistically
16 significant P-values rather than by focusing on missing studies or study size alone. The two most
17 commonly used approaches are the p-curve and p-uniform analyses, which are closely related but
18 serve distinct inferential purposes [19, 32].

19 The p-curve analysis is based on the idea that the distribution of statistically significant P-
20 values contains information on whether the reported findings reflect a genuine underlying effect or
21 are mainly the result of selective reporting. When a true non-zero effect exists and statistical models
22 are correctly specified, very small P-values should occur more often than P-values just below
23 conventional significance thresholds, producing a right-skewed p-curve (**Fig. 2A**). In contrast, when

1 the true effect is null and statistically significant results arise mainly through selective reporting,
2 analytical flexibility, or repeated testing, significant P-values tend to cluster just below the
3 significance cutoff, yielding a flat or left-skewed distribution (**Fig. 2B, 2C**). From this perspective,
4 the p-curve functions primarily as a diagnostic tool, evaluating whether a set of statistically significant
5 results contains evidential value rather than estimating the magnitude of the treatment effect [19, 33].

6 The p-uniform analysis builds on similar logic but additionally attempts to estimate a bias-
7 adjusted effect size. It assumes that, conditional on statistical significance, P-values should follow a
8 uniform distribution under the true effect. By modeling this distribution, p-uniform analyses seek to
9 infer the underlying effect size that best explains the observed set of significant P-values, thus
10 operating more explicitly as a model-based adjustment method, whereas the p-curve focuses on
11 detecting the presence or absence of an evidential value.

12 Both approaches are appealing because they directly target the selective reporting of
13 statistically significant results, rather than assuming that entire studies are missing. However, this
14 strength is also their main limitation, as P-value-based methods rely on several strong assumptions,
15 namely, that statistical tests are correctly specified, analyzed P-values are independent, and P-values
16 represent a well-defined and unbiased set of all relevant significant results [19, 33].

17 From an editorial and methodological perspective, p-curve results require careful
18 interpretation and may be unreliable in several common clinical research settings. The p-curve should
19 not be relied upon when P-values come from extensive outcome multiplicity, flexible analyses, or
20 undisclosed outcome switching, because the set of “significant” results may itself have been
21 selectively assembled. The method is also unreliable when statistical models are specified incorrectly,
22 multiple P-values from the same study are treated as independent, or study selection is driven by
23 effect direction or novelty rather than statistical significance. Under these conditions, a right-skewed

1 p-curve may reflect selective reporting or analytical flexibility rather than a true evidential value.
2 Therefore, p-curve results should not be interpreted in isolation, but rather be evaluated alongside
3 protocols, trial registries, risk-of-bias assessments, and other publication bias analyses.

4 However, these assumptions are often violated in clinical studies. Trials frequently evaluate
5 multiple outcomes, time points, subgroups, and analytical specifications, reporting only a subset of
6 favorable results. When such multiplicity, outcome switching, or selective analysis occurs, the
7 observed P-value distribution itself may be selectively constructed. The p-curve may thus suggest an
8 evidential value, even when the apparent signal is driven by selective reporting, and p-uniform
9 estimates may become unstable or misleading [8, 19].

10 Therefore, p-curve and p-uniform analysis results should not be interpreted in isolation but
11 are best used alongside information from study protocols, trial registries, selective outcome reporting
12 assessments, and other publication-bias diagnostics. Rather than providing definitive proof for or
13 against publication bias, P-value-based methods should be viewed as a component of a broader
14 sensitivity analysis framework, helping to clarify the extent to which conclusions depend on
15 assumptions regarding reporting practices and selection mechanisms [8, 19].

16 ***Detection as a transitional step***

17 Methods to detect publication bias, whether graphical or statistical, do not explain why a
18 pattern appears; they only show that the observed relationship between effect size and precision does
19 not fit a simple reference model [7, 14-16]. Treating detection as the final step (e.g., treating any
20 significant Egger's test as definitive proof of publication bias and immediately reporting a single
21 "corrected" estimate) encourages premature adjustment and overconfident conclusions [7, 14-16].
22 Detection is rather a transitional step in an integrated inferential framework. It should prompt a careful
23 search for plausible explanations, such as heterogeneity, risk-of-bias gradients, outcome reporting

1 bias, or selective analysis, and aid in the selection of sensitivity analyses and methods to adjust for
2 publication bias that are conceptually aligned with those explanations [7, 14-16]. Thus, detection acts
3 as a bridge between descriptive diagnostics and the counterfactual modeling and interpretive work
4 developed in the following sections [8, 9, 15].

5

6

Epub ahead of print

1 Statistical adjustment for publication bias

2 *Adjustment as counterfactual modeling*

3 Methods to adjust for publication bias are often described as solutions to publication bias,
4 but logically they do not “remove” the bias in a causal sense [12–15]. Instead, they encode explicit
5 or implicit assumptions about the selection process that might be generating the asymmetry and then
6 compute how the pooled effect would appear if those assumptions were true [3–8,12–15]. The
7 resulting adjusted estimates are therefore hypothetical and model-dependent. This dependence on
8 untestable assumptions is thus a central feature that should be made explicit rather than concealed
9 [12–15]. From a sensitivity-analysis perspective, the key question is not which adjusted estimate is
10 “correct,” but how much the substantive conclusions change or remain stable across a range of
11 plausible selection and reporting mechanisms [12–15].

12 *Trim-and-fill: symmetry by construction*

13 Trim-and-fill works by editing the funnel plot so that it looks symmetric. In practice, the
14 algorithm first “trims” small, extreme studies from one side of the funnel and estimates where the
15 center would be under symmetry. It then “fills” in mirror-image studies on the other side and
16 recomputes the pooled effect. This produces a clear visual story: “here are the missing negative
17 studies, and here is the corrected effect,” but the filled-in studies are not real data; they are values
18 created by the algorithm under specific assumptions [7, 8, 12, 13, 16]. As a result, the apparent
19 symmetry can be visually compelling yet methodologically misleading because graphical persuasion
20 may outpace evidential justification (**Fig. 3**).

21 This method assumes that (1) the asymmetry is mainly caused by missing small, unfavorable
22 studies, and (2) these missing studies would be arranged symmetrically around the pooled effect if
23 they could be observed. These assumptions are often unrealistic when substantial heterogeneity or

systematic differences in the study designs, outcome definitions, follow-up, or risk of bias exist. If the funnel is asymmetric because true effects differ across subgroups, outcomes are defined differently, or high-risk-of-bias studies tend to show larger effects, trim-and-fill can be misleading as it will “create” missing studies that reflect the geometry of the algorithm rather than any plausible selection process in the real research ecosystem [7, 8, 12, 13, 16].

In such settings, the adjusted pooled effect can reflect the enforced symmetry of the method more than the clinical reality. Empirical work and case studies show that trim-and-fill can add several moderately sized, near-null studies that do not resemble any observed trial and can either over-correct or under-correct depending on how heterogeneity and small-study effects interact. For example, when small high-quality trials show larger effects than larger low-quality trials, trim-and-fill may “correct” toward the larger, biased studies rather than toward the truth.

For these reasons, trim-and-fill is best used as a simple, transparent sensitivity analysis, not as a definitive fix for publication bias. It is most informative when framed explicitly as “If we assume that asymmetry is entirely due to missing small, unfavorable studies arranged symmetrically around the pooled effect, here is how our conclusions would change.” In an integrated framework, its role is to illustrate how sensitive the main conclusions are to this specific symmetry-based “missing-study” scenario, alongside other methods to adjust for publication bias that encode different assumptions about selection and reporting [7, 8, 12, 13, 16, 34].

Selection models: explicit selection mechanisms with strong sensitivity to specification

Selection models attempt to describe the likelihood that a study will be “seen” (e.g., published and then included in a meta-analysis) as a function of its P-value, effect direction, or other study characteristics. In a common simple version, the analyst specifies a selection function that assigns various publication probabilities to different *P*-value ranges (e.g., one probability for $P < 0.01$,

another for $P = 0.01-0.05$, and another for $P \geq 0.05$) and then jointly estimates the underlying effect and these selection parameters from the observed studies [8, 9, 15, 16]. This is attractive because it directly encodes the idea that “significant” or favorable results are more likely to appear in the literature, but it also relies on strong assumptions about the exact shape (cut-points, patterns) and strength (likelihood) of selection, which are difficult to identify reliably from a single meta-analysis with limited data [8, 9, 15, 16].

Crucially, the true selection process is unknown [6,12–15]. Different plausible selection functions (i.e., using alternative P-value cut-off points, different up-weighting of significant results, or asymmetric treatment of positive and negative findings) can often fit the observed data almost equally well while producing considerably different adjusted effect estimates [6,12–15]. Simulation studies and applied examples show that even modest changes in how strongly significant studies are up-weighted, or in how many “non-significant” studies are assumed to be suppressed, can move the adjusted pooled effect from clearly beneficial to nearly null, even when standard fit statistics and residual checks look similar [6,12–15]. In other words, selection models can be highly sensitive to specification as small changes in the assumed selection mechanism can lead to large changes in the “corrected” effect [6,12–15].

Therefore, selection models are best used as part of a structured sensitivity analysis rather than as a one-shot correction that yields a single preferred estimate. In practice, this means fitting several plausible selection models (for example, different P-value cut-off points or different assumptions about the extent that positive results are favored), comparing the extent to which the main clinical conclusions change across these models, and, where possible, anchoring the choice of selection functions in external information such as trial registries, protocols, reporting guidelines, or prior empirical evidence about editorial and reporting practices. Within an integrated inferential framework, the main value of selection models is not that they can recover the “true” effect, but that

they make the assumed selection process explicit and allow readers to see the extent to which the conclusions depend on those assumptions [8, 9, 15, 16].

PET–PEESE and related regression adjustments

Regression-based approaches such as the precision-effect test (PET) and precision-effect estimate with standard errors (PET–PEESE) are designed to address small-study effects by modeling how estimated treatment effects change with study precision, usually measured using the standard error or its variance. These methods are based on the simple idea that if smaller, less precise studies are more likely to report extreme results and larger, more precise studies are less affected by selective dissemination, then the relationship between effect size and precision may reveal the influence of small-study effects.

With the PET, effect sizes are regressed on the standard error, and the regression line is extrapolated to a standard error of zero. The intercept represents the expected effect in an infinitely precise study, and is interpreted as a bias-adjusted estimate. Importantly, an infinitely precise study does not correspond to any real clinical trial, rather, it is a conceptual reference point, a model-based projection that relies on strong assumptions about how effect size changes as precision increases (**Fig. 4A**). From an interpretive standpoint, the PET estimate is best understood as a test of whether a meaningful treatment effect remains after accounting for the association between the effect size and imprecision.

These methods are appealing because they are conceptually simple, easy to implement, and naturally fit within standard meta-regression frameworks. However, their validity depends on the key assumption that the relationship between effect size and precision is approximately linear and monotonic. This assumption is unlikely to be valid in several clinical meta-analyses. When heterogeneity is substantial, the precision range is dominated by a few influential studies, or study

precisions vary only within a narrow window, PET estimates can become unstable [35].

The PEESE approach is often applied after the PET and is interpreted as a refinement that estimates the likely magnitude of the treatment effect under the assumption that a true effect exists but is distorted by small-study effects. The PEESE models the effect size as a function of variance (squared standard error) rather than the standard error itself (**Fig. 4B**). This alternative specification can lead to different adjusted estimates, as the divergence between the PET and PEESE regression lines illustrates (**Fig. 4C**). The PET and PET-PEESE should be interpreted together, as the PET primarily addresses whether the observed effect survives adjustment for imprecision, whereas the PET-PEESE suggests the possible size of that effect under the assumed precision-effect relationship.

In meta-analyses that include a few large pragmatic trials along with several smaller explanatory trials, the PET-PEESE may effectively shift the synthesis toward the largest and most precise studies. Rather than “correcting” bias in a general sense, this method may mainly reflect the results of these large trials. **Fig. 4D** illustrates this pattern by plotting risk ratios against study precision: smaller, less precise studies tend to report larger benefits, whereas the largest and most precise trials report effects closer to the null. This pattern highlights both the motivation for the PET-PEESE and its limitations, as it performs well only if the observed precision-effect relationship truly reflects small-study bias rather than genuine differences between trials.

Simulation studies show that the PET-PEESE can perform reasonably well under certain selection mechanisms and low heterogeneity but may perform poorly when heterogeneity is moderate to high or when selection operates through mechanisms other than simple significance filtering. For these reasons, the PET-PEESE should be viewed as one element of a broader sensitivity-analysis toolkit and not as a standalone method to adjust for publication bias [14, 36].

Bayesian approaches to publication bias adjustment

Bayesian approaches to publication bias provide a flexible framework for modeling selective reporting by treating both the data-generating process and selection mechanism as probabilistic entities. Rather than producing a single bias-corrected estimate through algorithmic adjustment, Bayesian models explicitly encode assumptions about study selection, heterogeneity, and prior beliefs about effect sizes within a unified inferential structure. This allows uncertainty arising from missingness, small-study effects, and between-study variability to propagate through the posterior distribution rather than being handled in a piecemeal fashion [37-39].

In practice, Bayesian publication bias models often specify a likelihood for observed study results together with a selection function that governs the probability of a study being observed or published, conditional on its statistical significance or effect size. Priors are then placed on both the treatment effect and the parameters of the selection process. This formulation makes explicit what is implicit in many frequentist adjustments, as conclusions depend critically on assumptions about how studies enter the literature and how strongly selection depends on statistical results. Importantly, these assumptions are not resolved by data alone and must be justified substantively rather than statistically [37-39].

Transparency about assumption dependence is one conceptual advantage of Bayesian approaches. By varying priors or selection models, analysts can directly examine changes in posterior estimates under alternative, but plausible, mechanisms of selective reporting. Disagreement between posterior summaries across models is, therefore, informative, signaling the fragility of the evidence rather than the failure of the method. In this sense, Bayesian approaches align naturally with the view of publication bias adjustment as a structured sensitivity analysis rather than as a causal correction [15, 37, 38].

However, Bayesian models are equally subject to the fundamental limitations of publication

1 bias adjustment. When the number of studies is small, heterogeneity is substantial, or selection
2 mechanisms are complex and poorly understood, posterior inferences can be largely driven by prior
3 choices. Highly informative priors may stabilize estimates but risk masking uncertainty, whereas
4 weakly informative priors may yield diffuse posteriors that offer limited practical guidance. Moreover,
5 different selection models can fit the observed data equally well while implying markedly different
6 conclusions, underscoring the non-identifiability of publication bias mechanisms from the published
7 literature alone [37-39].

8 Accordingly, Bayesian publication bias methods are best used to make uncertainty explicit
9 rather than to eliminate it. Their primary contribution lies in clarifying how conclusions depend on
10 assumptions about selection, heterogeneity, and prior beliefs and in facilitating coherent comparisons
11 across competing inferential scenarios. When reported transparently, along with frequentist
12 adjustments, risk-of-bias assessments, and exploration of heterogeneity, Bayesian approaches can
13 strengthen interpretive discipline without creating an illusion of definitive bias correction [37-39].

14 In practice, Bayesian publication bias models are most commonly formulated as selection,
15 weight function, or hierarchical models with explicit selection components, differing primarily in
16 how the probability of study inclusion is linked to statistical significance, effect size, or both [37-39].
17 Importantly, these Bayesian formulations do not obviate the need for a sensitivity analysis; rather,
18 they formalize it by embedding selection assumptions directly into the inferential model.

19 ***Adjustment as sensitivity analysis: what to report***

20 Across all adjustment methods, including both the frequentist and Bayesian approaches,
21 disagreements between results should be viewed as informative rather than indications of failure.
22 Such divergence shows that the main conclusions depend on assumptions about study selection and
23 reporting, which cannot be verified directly [4, 8, 15, 16]. When different methods such as trim-and-

1 fill, selection models, PET–PEESE, and P-value-based approaches produce clearly different adjusted
2 effects, this signals assumption dependence, not analytical error.

3 Therefore, the most appropriate response is transparency. Analysts should clearly state the
4 assumptions each method relies on, the sensitivity of the key clinical conclusions to those assumptions,
5 and how this fragility may affect the certainty of the evidence and strength of recommendations [4,
6 6, 8, 15, 16]. Instead of presenting a single “corrected” estimate, meta-analyses should report a
7 structured set of sensitivity analyses. These may include worst-case bias scenarios and ranges of
8 estimates derived under alternative assumptions, for example through different priors or selection
9 models in a Bayesian framework [4, 8, 15, 16]. The goal is to demonstrate which effect sizes remain
10 compatible with the data under plausible selection mechanisms.

11 From this perspective, adjustment methods including Bayesian selection models are not
12 automatic correction tools. Rather, they are tools for disciplined inference that help define the limits
13 of what can reasonably be concluded [12–15]. When reported transparently and interpreted alongside
14 risk-of-bias assessments, exploration of heterogeneity, and consideration of information size, these
15 methods help align clinical conclusions with the true robustness of the evidence [4, 6, 8, 15, 16].

16 In settings where heterogeneity is low, outcomes and time points are clearly prespecified, and
17 external information (e.g., trial registries) constrains plausible selection processes, publication bias
18 adjustment methods can be particularly informative as sensitivity analyses. Under these conditions,
19 the Bayesian and frequentist approaches may show greater convergence, supporting more confident
20 but conditional interpretations. However, their role remains inferential rather than corrective.

1 **An integrated inferential framework**

2 Publication bias analysis is best viewed as a repeating cycle of reasoning rather than a
3 checklist of separate techniques [8, 9, 15, 16]. In this cycle, the detection methods highlight where
4 the data deviate from a simple reference model. Adjustment methods then offer structured
5 explanations for these deviations and generate alternative effect estimates under explicit assumptions
6 regarding selection. Both frequentist and Bayesian approaches can be used at this stage, with
7 interpretations demonstrating how clinical conclusions should change in light of these competing
8 scenarios [2, 7-9, 14-16, 40]. The goal is not to identify a single “best” method, but to understand
9 how strongly the direction, size, and certainty of conclusions depend on assumptions about which
10 studies and results are more likely to appear [8, 9, 15, 16].

11 In this framework, the first step is detection. Graphical and statistical diagnostics (e.g., funnel
12 plots, Egger’s and Begg’s tests, P-value-based approaches, and other measures of small-study effects)
13 are used to stress test a simple model with homogeneous effects and no selection. Detection results
14 are interpreted as signals that the simple model may be strained and not as proof of publication bias
15 or any specific mechanism. Their main role is to prompt consideration of alternative explanations,
16 including between-study heterogeneity, risk-of-bias gradients, outcome reporting bias, selective
17 analysis, and design-related differences [6-8, 16, 19].

18 The second step is model-based adjustment, in which methods such as trim-and-fill, selection
19 models, PET–PEESE, and Bayesian selection or hierarchical models are applied as sensitivity
20 analyses under different narratives of how selection operates [7, 8, 12-15, 32, 35]. Each method
21 encodes specific assumptions: trim-and-fill assumes symmetry around a pooled effect; selection
22 models assume explicit, often P-value-dependent publication probabilities; regression-based
23 approaches assume a defined relationship between effect size and precision; and Bayesian models

embed these assumptions probabilistically through priors on effects, heterogeneity, and selection mechanisms [7, 8, 12-15, 32, 35]. Rather than concealing these assumptions, the framework treats them as central objects of inference and encourages analysts to explore a range of plausible models, documenting how pooled estimates and uncertainty change across this range [8, 9, 16].

The third step is interpretation, in which the results from detection and adjustment are combined into judgments regarding robustness and confidence. The key questions include the following: How wide is the range of effect sizes compatible with the data under plausible selection mechanisms? Do conclusions regarding benefits, harm, or equivalence remain stable across methods, including Bayesian and frequentist models, or do they hinge on specific assumptions? When these findings are considered alongside risk of bias assessments, heterogeneity, and limitations in information size, do they justify downgrading the certainty of the evidence or softening clinical recommendations? [6, 8, 9, 16, 18]. By centering these questions, the framework shifts attention away from selecting a single “corrected” estimate and toward calibrating claims in proportion to evidential fragility or stability.

In practical terms, the integrated framework operates as a cycle: (1) specify a baseline meta-analytic model and apply detection tools; (2) define plausible selection and reporting mechanisms and apply corresponding adjustment methods as sensitivity analyses, including alternative Bayesian priors or selection structures when appropriate; and (3) compare the resulting estimates and intervals, explicitly mapping how conclusions depend on assumptions and data limitations [6, 8, 9, 16, 18]. The worked example, together with later discussions of common misuses, editorial evaluations, and umbrella reviews, illustrates how this inferential loop can guide both authors’ analytical choices and editors’ judgments on whether publication bias has been handled with appropriate rigor and restraint [6, 8, 9, 16, 18].

1 **Applied example in clinical meta-analysis**

2 Consider a hypothetical random-effects meta-analysis of ten randomized controlled trials
3 evaluating tropisetron versus a control for prevention of postoperative nausea and vomiting (PONV)
4 in adult patients undergoing surgery under general anesthesia. The primary outcome is the incidence
5 of PONV across the postoperative period and in the event that data were reported at multiple time
6 points, the first available postoperative time point is selected as the outcome of interest.

7 A conventional random-effects model yields a pooled risk ratio (RR) of 0.678 (95% CI:
8 0.561–0.821), which would typically be interpreted as evidence of a clinically meaningful benefit.
9 However, the evidence base is modest and heterogeneous ($I^2 = 64/5\%$, prediction interval: 0.40–1.14)
10 (**Fig. 5**). The total sample size is limited ($n = 1502$), event rates vary widely (from 12.0% to 81.6%),
11 outcome timing is inconsistent (from 0 to 18 h postoperatively), and study designs differ
12 substantially between small explanatory trials and larger pragmatic trials. These features create a
13 setting in which small-study effects are plausible, even in the absence of selective publications.

14 A funnel plot shows an apparent lack of small, unfavorable trials and a cluster of small
15 studies with large beneficial effects (**Fig. 1A**). Egger's regression test yields a P-value of 0.001, and
16 Begg's rank correlation test is borderline ($P = 0.089$), which might be interpreted as evidence of
17 publication bias. However, closer inspection suggests that small trials differ systematically from
18 large trials in terms of methodological quality (**Fig. 1B**). In this context, funnel plot asymmetry may
19 reflect effect modification or methodological gradients rather than selective non-publication alone
20 [7, 12, 13, 41].

21 To explore this uncertainty, publication bias analyses were treated as structured sensitivity
22 analyses rather than corrective procedures [12, 13]. Trim-and-fill attenuated the pooled effect, with
23 the adjusted CI crossing the null (from RR 0.678, 95% CI: 0.561–0.821 to RR 0.827, 95% CI: 0.679–

1 1.007); however, the imputed studies did not resemble any clearly identifiable group of real trials
2 (**Fig. 3**). This highlights the fact that the adjustment reflects the symmetry assumption of the method
3 rather than a validated model of the publication process [7, 12, 13, 41].

4 Selection models provide a complementary perspective on potential publication bias. Three
5 P-value-based selection scenarios were examined to assess the sensitivity to increasingly strong and
6 granular assumptions regarding selective publication. In Scenario 1, studies with $P < 0.05$ were
7 assumed to be twice as likely to be observed as non-significant studies, yielding a modestly
8 attenuated pooled estimate (RR 0.681, 95% CI: 0.545–0.851). Scenario 2 imposed a stronger
9 selection assumption, a three-fold preferential observation of statistically significant studies, yet
10 produced an almost identical estimate (RR 0.681, 95% CI: 0.545–0.851). Scenario 3 applied a
11 gradient with greater differentiation ($P < 0.01$: 2×; $0.01 \leq P < 0.05$: 1.5×; $P \geq 0.05$: reference), again
12 resulting in virtually unchanged pooled effects (RR 0.680, 95% CI: 0.544–0.850) (**Table 2**).

13 Importantly, all selection-adjusted estimates were very close to the original random-effects
14 result (RR 0.678, 95% CI: 0.561–0.821), indicating that P-value-based selection assumptions (weak
15 or strong) had a limited influence on the pooled effect in this evidence base [8, 9, 15, 16].

16 However, regression-based approaches (PET–PEESE) yielded markedly different
17 conclusions. The PET produced an estimate close to the null (RR = 1.018, 95% CI: 0.945–1.096),
18 while the PEESE suggested an intermediate effect (RR = 0.889, 95% CI: 0.817–0.966) (**Fig. 4A–C**).
19 These differences arose because each method places implicit trust in different regions of the evidence
20 base (i.e., large pragmatic trials, small explanatory trials, or a compromise between them) (**Fig. 4D**).
21 The divergence is, therefore, informative, as it shows that conclusions depend on judgments
22 regarding which studies are most credible [35].

23 Whereas the selection models explored above adopt a frequentist framework, Bayesian

approaches offer an alternative inferential structure that makes explicit assumptions about selection, heterogeneity, and prior beliefs within a unified probabilistic model [37-39]. To illustrate how Bayesian methods clarify assumption dependence, we applied a series of Bayesian models to a tropisetron meta-analysis with varying prior specifications and selection mechanisms. A baseline Bayesian model using a weakly informative prior (normal [0, 1] on the log risk ratio scale) yielded a posterior mean RR of 0.681 (95% credible interval [CrI]: 0.563–0.823), nearly identical to the frequentist random-effects estimate (RR 0.678, 95% CI: 0.561–0.821), demonstrating that Bayesian and frequentist approaches converge when selection bias is not explicitly modeled and priors are diffuse (**Fig. 6, Table 2**).

After selection mechanisms were incorporated into the Bayesian framework, the posterior estimates shifted in a systematic and interpretable manner. Under Scenario 1, which assumed a two-step selection process in which studies with $P < 0.05$ were twice as likely to be published, the adjusted posterior estimate was RR 0.631 (95% CrI: 0.537–0.740). Imposing a stronger selection gradient in Scenario 2 (i.e., a threefold preferential observation of statistically significant studies) shifted the posterior toward smaller relative risk values, yielding an RR of 0.604 (95% CrI: 0.525–0.696). Scenario 3, which applied a moderate and more differentiated gradient ($P < 0.01$: 2×; $0.01 \leq P < 0.05$: 1.5×), produced an intermediate estimate (RR 0.637, 95% CrI: 0.541–0.750) (**Fig. 6, Table 2**).

In these scenarios, stronger assumptions regarding selective publication were associated with numerically larger estimated treatment effects. Under these assumptions, studies with null or modest effects contribute less to the posterior inferences, while those reporting larger benefits carry greater weight. Consequently, the estimated effect increases. This behavior reflects how the model distributes evidence under the assumed selection mechanisms and does not prove the presence of publication bias.

Prior specifications influenced the posterior inferences, although their impact was smaller than that of the selection mechanism. Table 2 summarizes a prior sensitivity analysis across four specifications that reflect progressively stronger prior skepticism about treatment benefits: weakly informative (normal [0, 3.16]), moderately informative (normal [0, 1]), strongly informative (normal [0, 0.447]), and very skeptical (normal [0, 0.316]).

Under the very weak and moderate priors, both of which place minimal constraints on the effect size and allow the data to dominate, the posterior estimates closely matched the original random-effects meta-analysis (RR 0.678, 95% CI: 0.561–0.821), yielding an RR of 0.679 (95% CrI: 0.561–0.821) and 0.681 (95% CrI: 0.563–0.823). This indicates that the Bayesian analysis largely reproduces the frequentist result in the absence of strong prior beliefs.

As prior skepticism increased, the posterior estimates gradually shifted toward the null. The skeptical prior produced an intermediate estimate (RR 0.690, 95% CrI: 0.573–0.832), reflecting a balanced position between prior belief and observed data. Under the very skeptical prior, which encodes a strong belief that the true effect is likely to be small or absent, the posterior moved further toward the null (RR 0.702, 95% CrI: 0.585–0.842).

Critically, the Bayesian framework makes this assumption dependence explicit and quantifiable, whereas it can be implicit in frequentist adjustments. By reporting posterior estimates under multiple priors and selection models side by side, analysts acknowledge that inferential conclusions are conditional on model choice and invite readers to assess which assumptions seem most plausible given substantive knowledge of the field. The resulting synthesis systematically and transparently documents uncertainty rather than eliminating it.

When analyses are restricted to statistically significant trial-level results ($P < 0.05$), p-curve methods indicate a right-skewed distribution of P-values, which is consistent with the evidential

value under standard assumptions. As seen in **Fig. 7A**, among the five studies with statistically significant results, the P-values cluster toward smaller values (left side of the histogram, 0.00–0.01 range), producing the characteristic right-skewed distribution that the p-curve interprets as evidence against null findings being driven purely by selective reporting.

However, conducting a comparison with trial registry information reveals outcome switching and multiplicity in several studies, indicating that the observed set of significant P-values is itself selectively constructed. **Fig. 7B** illustrates this critical limitation by displaying each study's significance level along with evidence of reporting irregularities. Four of the five significant studies (Studies B, F, J, and D) show registry discrepancies: two with outcome switching (orange bars) and two with multiplicity (darker orange bars). In contrast, the five non-significant studies ($P \geq 0.05$) show no evidence of registry discrepancies, all classified as "None" (blue bars).

This pattern demonstrates that even a superficially reassuring p-curve can reflect a selectively constructed evidence base, rather than robust evidential value. The concentration of reporting irregularities, specifically among significant studies (80% of significant studies vs. 0% of non-significant studies), suggests that the right-skewed p-curve may itself be an artifact of selective outcome reporting rather than evidence of true treatment effects. This limits the reassurance that P-value-based methods can provide in this context and underscores the fact that apparent evidential value does not necessarily imply robustness when selective outcome reporting is present.

Fig. 6 presents a comprehensive comparison of frequentist (blue) and Bayesian (red) approaches to publication bias adjustment, illustrating the assumption dependence of meta-analytic inference when small-study effects are present. Frequentist methods yield divergent estimates depending on their underlying assumptions: trim-and-fill approaches the null ($RR = 0.827$, 95% CI: 0.679–1.007) by imputing five missing studies, selection models produce estimates of $RR = 0.680$ –

0.681 under varying selection severity assumptions, and regression-based methods range from no effect (PET: $RR = 1.018$, 95% CI: 0.945–1.096) to moderate benefit (PEESE: $RR = 0.889$, 95% CI: 0.817–0.966). Bayesian approaches show similar point estimates ($RR = 0.604$ – 0.702), with tighter credible intervals but meaningful sensitivity to prior specification, shifting from $RR = 0.679$ (very weak prior) to $RR = 0.702$ (very skeptical prior). Therefore, substantial divergence across methods or prior specifications should be interpreted as a signal that clinical conclusions are sensitive to modeling assumptions and warrant cautious interpretation.

The central inferential message is not that one adjustment method “wins,” but rather that the claim of benefit is fragile across plausible assumptions about selection mechanisms and prior beliefs [4, 8, 16]. The divergence of adjusted estimates, spanning from substantial benefit ($RR = 0.604$) to no effect ($RR = 1.02$), combined with evidence of selective outcome reporting in 80% of significant studies (**Fig. 7B**), indicates that clinical recommendations should be tentative with moderate or low certainty ratings. Rather than selecting a single “best” estimate, decision-makers should consider the entire range of plausible effect sizes when evaluating treatment efficacy, treating the divergence across methods as a result that reflects epistemic uncertainty rather than a methodological problem to be solved [4, 6, 8, 16].

1 **Common misuses in clinical meta-analyses**

2 Empirical assessments of meta-analyses in clinical medicine consistently show that
3 publication bias procedures are often applied in a mechanical, checklist-driven way, with little
4 attention to their assumptions or limits of interpretation [25, 42, 43]. One recurring error is the
5 tendency to treat funnel plot asymmetry or a statistically significant Egger's test as direct proof of
6 publication bias, even when substantial heterogeneity, inconsistent outcome definitions, or clear
7 design-related trends across studies exist. Under such conditions, the observed asymmetry can
8 reasonably be explained by real-effect modification, differences in study quality, baseline risk, or
9 follow-up duration, especially when these factors are related to the study size. Calling all such
10 patterns "publication bias" oversimplifies the problem and ignores other plausible explanations,
11 collapsing multiple mechanisms into a single narrow label [2, 6, 18, 19].

12 A second common misuse is the presentation of one bias-adjusted estimate, most often from
13 a trim-and-fill or single selection model, as the main or final result of the meta-analysis. In many
14 studies, the unadjusted analysis or results from alternative models are pushed into supplementary
15 files or not shown [8, 9, 12, 13, 15, 16, 44]. This turns a sensitivity analysis into a headline conclusion,
16 although different and equally reasonable models can fit the same data and lead to clearly different
17 pooled effects [8, 9, 14-16]. A related problem is that researchers rarely carefully examine how and
18 why the estimate changes after adjustment. Adjusted results are often assumed to be more accurate
19 simply because they are smaller or closer to no effect without considering whether the studies that
20 drive this change make clinical or methodological sense.

21 Third, measures of heterogeneity such as I^2 , τ^2 , τ , or prediction interval are frequently
22 reported but not meaningfully used in interpretation [3]. Authors may acknowledge large between-
23 study variability yet still apply asymmetry tests and bias-adjustment methods as if all studies were

1 estimating the same underlying effect. This internal inconsistency weakens both the detection of bias
2 and any subsequent adjustments because most publication bias methods rely implicitly on relatively
3 homogeneous effects [3, 16].

4 Finally, and perhaps most consequentially, overconfident conclusions remain commonplace.
5 Even when publication bias analyses show instability, such as widely varying adjusted estimates,
6 broad sensitivity ranges, or inadequate information size, authors may still use strong causal language
7 or recommend clinical adoption without appropriately lowering certainty [5, 6, 8, 9, 14-16].
8 Collectively, these practices reflect a persistent misunderstanding of publication bias adjustment
9 methods as corrective tools that “fix” biased evidence, rather than as exploratory tools that test the
10 extent to which conclusions depend on untestable assumptions about study selection and reporting
11 [8, 9, 14-16].

1 Editorial and authorial evaluation: what constitutes adequate handling of publication bias?

2 From an editorial and methodological perspective, handling publication bias in meta-
3 analyses in clinical medicine cannot be judged simply by whether a funnel plot or Egger's test has
4 been reported [5, 6, 16, 18, 25]. The same limitation applies from an authorial perspective: adequate
5 handling is not demonstrated by performing these analyses, but by their framing, justification, and
6 interpretation. The key question should be framed as "Were these tools used thoughtfully?" Critically,
7 the manuscript should show inferential discipline, with authors recognizing that small-study effects
8 are non-specific, discussing other plausible explanations besides publication bias, and clearly
9 explaining why the chosen sensitivity analyses are appropriate for this particular body of evidence
10 [7-9, 15, 17, 25]. For authors, reviewers, and editors alike, this means checklist-style evaluations are
11 avoided, and both the internal logic and coherence of the analysis are judged. In particular, attention
12 should be paid to how publication bias assessments are integrated with risk-of-bias evaluations,
13 explorations of heterogeneity, checks for selective outcome reporting, and judgments about the
14 certainty of the evidence, rather than being treated as isolated statistical steps [5, 6, 8, 9, 15, 18].

15 From this shared perspective, a role-informed checklist (**Table 3**) can help clarify
16 expectations across authorship, peer review, and editorial decision making. Rather than functioning
17 as a procedural requirement, the checklist is intended for structural interpretation. The focus is on the
18 framing, justification, and integration of publication bias analyses into the overall inferential narrative
19 and not simply on whether particular statistical techniques were applied [4, 8, 12, 13, 15, 17]. In this
20 sense, the checklist serves as a guide for judging studies rather than as a substitute. By foregrounding
21 questions about assumptions, coherence, and evidential implications, it aims to promote consistency
22 in evaluations, while discouraging the mechanical use of bias diagnostics as standalone indicators of
23 rigor.

The interpretive role of publication bias analyses aligns closely with the logic of evidence certainty frameworks, such as GRADE [45]. Rather than serving as a mechanism to correct effect estimates, publication bias assessments should inform judgments regarding the credibility and stability of the evidence [46]. When adjusted estimates vary substantially across plausible models or when conclusions depend strongly on unverifiable selection assumptions, this sensitivity should be reflected as a downgrade in certainty rather than resolved through the selection of a single preferred estimate. Therefore, from both authorial and editorial perspectives, the appropriate response to suspected publication bias is not statistical correction, but transparent acknowledgment of uncertainty and proportionate calibration of evidentiary claims.

Editors and reviewers can promote better practice by valuing manuscripts that use publication bias methods to explore the robustness of conclusions, rather than to produce a single “corrected” effect estimate [4, 8, 15, 17]. Authors should frame bias adjustment explicitly as a sensitivity exercise and not as a substitute for the primary analysis. Well-handled analyses treat bias adjustment as a sensitivity exercise, and not as a replacement for the primary analysis. This includes encouraging the regular use of structured sensitivity analyses (e.g., worst-case bias scenarios, model-averaging approaches, and clear demonstrations of the change in the conclusions under different selection assumptions) and explicitly downgrading the certainty of evidence when conclusions depend strongly on unverifiable assumptions [4-6, 8, 15, 17]. Ultimately, treating publication bias analyses as part of the interpretation rather than as a final diagnostic checkbox can help journals ensure that synthesized evidence is both methodologically rigorous and appropriately cautious in its claims [4-6, 8, 15, 17].

Preprints and the reshaping of the publication bias ecosystem

The availability of preprints is currently changing the channels of evidence sharing in clinical

1 medicine that impact publication bias and meta-analyses. Preprints are often described as a solution
2 to publication bias because they allow for the rapid public posting of study results without the need
3 for journal gatekeeping. This can reduce outright non-publication and time-lag bias because negative
4 or inconclusive studies can, in principle, be posted as easily as positive ones. However, preprints do
5 not remove selective processes; instead, they change when, how, and which results become visible,
6 thereby introducing new forms of selection while mitigating others [1, 47, 48].

7 Traditional publication bias analysis methods were developed for a journal-centered world,
8 which implicitly assumes a relatively static evidence base in which studies are either published or not
9 published. Preprints add an additional dissemination layer that is both dynamic and versioned.
10 Decisions about whether to post a preprint, which analyses to present, and whether or how to update
11 later versions remain under the control of the investigators. As a result, preprints change what is
12 missing and when rather than eliminating missingness altogether [1, 48].

13 This shift has important implications for statistical signals of publication bias. Funnel plot
14 asymmetry and small-study effects may partly reflect the temporal ordering of evidence. Early
15 preprints may come mainly from small, exploratory studies with imprecise and sometimes
16 exaggerated effects, with larger, more definitive trials appearing only later. In such situations,
17 asymmetry does not necessarily indicate selective suppression, but may instead arise from the timing
18 and evolution of dissemination [1, 47].

19 For evidence synthesis, neither automatic inclusion nor systematic exclusion of preprints is
20 methodologically neutral. Including preprints may increase the heterogeneity and instability if many
21 are small, preliminary, or later retracted or heavily revised, whereas excluding them may reintroduce
22 a directionally biased synthesis by preferentially omitting null or unpublished results. A more
23 defensible strategy is to identify preprints transparently, track version histories, and perform

1 sensitivity analyses that compare meta-analysis results with and without preprint data [1, 47].

2 Preprints also interact with higher-order evidence synthesis in distinct ways. Umbrella
3 reviews tend to amplify bias vertically by stacking one synthesis on top of another, whereas preprints
4 redistribute bias horizontally across dissemination stages and over time. Recognizing this distinction
5 helps in understanding how publication and reporting biases propagate in modern evidence
6 ecosystems [1, 49].

7 Consistent with the broader argument of this review, preprints should not be treated as a cure
8 for publication bias. Instead, they represent a structural change in the research ecosystem that alters
9 the form, timing, and visibility of selective processes. In this setting, publication bias analysis remains
10 an exercise in mapping assumption dependence and evidential fragility, rather than in recovering a
11 single bias-corrected effect [1, 50].

1 **Amplification of publication bias in umbrella reviews**

2 Umbrella reviews synthesize evidence from multiple systematic reviews and meta-analyses,
3 and therefore operate one level above primary-study selection processes [5, 6, 9]. This structure
4 creates opportunities for publication bias to be amplified across synthesis layers, as bias that
5 originates in the primary literature (selective publication and reporting) is propagated into individual
6 meta-analyses (selective inclusion, analytic flexibility), and then further reinforced in umbrella
7 reviews that preferentially include prominent or statistically significant meta-analyses [5, 6, 9]. The
8 overlap of primary studies across meta-analyses can cause the same influential positive trials to be
9 counted repeatedly, creating a false impression of replication when the same selectively disseminated
10 evidence is reused across syntheses. This recycling effect is particularly problematic when a small
11 number of early or high-impact trials dominate multiple systematic reviews and meta-analyses.

12 Furthermore, umbrella reviews often classify evidence strength using simple decision rules
13 such as significance thresholds, effect size cut-offs, or criteria based on the number and size of
14 contributing studies [16–18]. These rules can be misleading when the pooled estimates themselves
15 are unstable. When evidence grading relies on results that are fragile, highly sensitive to publication
16 bias adjustments, strongly heterogeneous, or based on limited information size, umbrella reviews may
17 turn weak or uncertain signals into categories that appear authoritative and definitive [4-6, 8, 15, 17].
18 In this way, statistical uncertainty can be hidden behind formal labels of “strong” or “convincing”
19 evidence.

20 Publication bias assessment in umbrella reviews should therefore focus on patterns of
21 fragility, study overlap, and robustness across competing assumptions, rather than relying only on
22 single corrected estimates from individual systematic reviews and meta-analyses or uncritical
23 evidence grading systems [4-6, 8, 15, 17]. Approaches that explicitly address the overlap of primary

1 studies, compare results across alternative inclusion criteria and modeling choices, and summarize
2 publication bias signals across constituent meta-analyses are particularly important at this higher level
3 of synthesis [5, 6, 9]. Without such safeguards, umbrella reviews risk amplifying bias while providing
4 an illusion of increased certainty.

5

6

Epub ahead of print

1 **Future directions: automation, artificial intelligence, and the new face of bias**

2 Automation and artificial intelligence (AI)-assisted evidence synthesis introduce new
3 avenues for bias, beyond classical statistical selection mechanisms. Algorithms trained mainly on
4 published and highly cited studies tend to inherit the publication and citation biases already present
5 in these sources. Consequently, automated screening, ranking, or prioritization systems may
6 systematically favor large, positive, or high-profile studies, while smaller or null studies receive less
7 attention. This process can quietly reinforce the same biases that systematic review and meta-analysis
8 methods are meant to examine. Large language models and other generative tools add an additional
9 layer of risk. They can turn uncertain, assumption-dependent meta-analysis results into fluent,
10 convincing narratives, often downplaying uncertainty, masking fragility, or omitting key caveats
11 regarding publication bias, heterogeneity, or limited information size. When results are presented in
12 a clear and polished manner, readers may confuse how convincing they sound with how strong the
13 underlying evidence is [51].

14 Addressing these challenges requires methodological and governance safeguards in addition
15 to improved algorithms. Traditional publication bias analyses must be actively integrated into the
16 design and oversight of automated systems [4, 8, 15, 17]. Bias-aware retrieval strategies, such as
17 systematic searches of trial registries, forward and backward citation chasing, and the inclusion of
18 preprints and regulatory documents, are essential to counter the default tendency of automated tools
19 to focus on highly cited journal articles [5, 6, 18]. Without these countermeasures, automation risks
20 narrowing rather than broadening the evidence base.

21 In addition, the routine comparison of registry records with published reports, transparent
22 audit trails for automated screening and inclusion decisions, and human-in-the-loop interpretations
23 that explicitly consider publication and citation bias should become standard components of AI-

1 assisted evidence synthesis workflows [5, 6, 18]. Human oversight is particularly important at
2 interpretive stages, during which judgments about certainty and clinical relevance are made.

3 In this emerging landscape, publication bias is no longer only a statistical problem, but also
4 a socio-technical problem shaped by how evidence is retrieved, summarized, and communicated.
5 Aligning tool development, reporting standards, and editorial expectations is therefore essential to
6 ensure both that automation supports methodological rigor rather than amplifying bias and that
7 conclusions remain appropriately cautious and well calibrated to the underlying evidence [4-6, 8, 15,
8 17].

Epub ahead of print

1 **Conclusion**

2 Publication bias analysis methods do not causally correct for bias. Rather than recovering a
3 hidden “true” effect, they reveal the extent to which clinical conclusions depend on unverifiable
4 assumptions about study selection, reporting, and dissemination and the potential fragility of an
5 evidence base under plausible mechanisms of missingness and selective reporting. As such, their
6 primary value lies not in numerical adjustment, but in clarifying the limits of inference. Viewed
7 through a certainty-of-evidence framework such as GRADE, the primary contribution of publication
8 bias analysis is to guide the appropriate downgrading of confidence in the evidence. Divergent
9 adjusted estimates, wide sensitivity ranges, or strong dependence on unverifiable selection
10 assumptions signal fragility that should temper causal language and clinical recommendations and
11 should not be resolved by choosing a single preferred model.

12 In this sense, publication bias analysis functions as a bridge between statistical modeling and
13 responsible interpretation. When used to characterize uncertainty through structured sensitivity
14 analyses and transparent reporting, it supports interpretive discipline rather than false precision.
15 Treated in this way, publication bias analysis methods help ensure that conclusions drawn from
16 systematic reviews and meta-analyses, and the clinical recommendations that follow, remain
17 proportional to the strength and stability of the underlying evidence rather than being overstated by
18 apparent numerical certainty.

19
20 **Data Availability:** All data generated or analyzed during this study are included in this published
21 article and its supplementary information files.

References

1. Kang H. Beyond the paywall: the role of preprints in overcoming publication bias. *J Evid-Based Pract* 2025; 1: 7-11.
2. Almalik O. Adjusting for publication bias in meta-analysis with continuous outcomes: a comparative study. *Mathematics* 2025; 13: 3487.
3. Choi GJ, Kang H. Heterogeneity in meta-analyses: an unavoidable challenge worth exploring. *Korean J Anesthesiol* 2025; 78: 301-14.
4. Bartoš F, Maier M, Wagenmakers EJ, Nippold F, Doucouliagos H, Ioannidis JPA, et al. Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Res Synth Methods* 2024; 15: 500-11.
5. Hennessy EA, Johnson BT, Keenan C. Best practice guidelines and essential methodological steps to conduct rigorous and systematic meta-reviews. *Appl Psychol Health Well Being* 2019; 11: 353-81.
6. Johnson BT, Hennessy EA. Systematic reviews and meta-analyses in the health sciences: best practice methods for research syntheses. *Soc Sci Med* 2019; 233: 237-51.
7. Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol* 2009; 9: 2.
8. Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. *J R Stat Soc Ser C Appl Stat* 2020; 69: 1091-119.
9. Bartoš F, Maier M, Wagenmakers EJ, Doucouliagos H, Stanley TD. Robust Bayesian meta-analysis: model-averaging across complementary publication bias adjustment methods. *Res Synth Methods* 2023; 14: 99-116.
10. Nakagawa S, Yang Y, Macartney EL, Spake R, Lagisz M. Quantitative evidence synthesis: a

practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environ Evid* 2023; 12: 8.

11. Lee S. Heterogeneity in meta-analysis: a path toward more meaningful clinical evidence. *Korean J Anesthesiol* 2025; 78: 297-8.
12. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2004; 56: 455-63.
13. Duval S, Tweedie R. A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *J Am Stat Assoc* 2000; 95: 89-98.
14. Stanley TD. Limitations of PET-PEESE and other meta-analysis methods. *Soc Psychol Personal Sci* 2017; 8: 581-91.
15. Maier M, VanderWeele TJ, Mathur MB. Using selection models to assess sensitivity to publication bias: a tutorial and call for more routine use. *Campbell Syst Rev* 2022; 18: e1256.
16. Lin L, Chu H. Quantifying publication bias in meta-analysis. *Biometrics* 2018; 74: 785-94.
17. Lin L, Chu H. Rejoinder to “quantifying publication bias in meta-analysis”. *Biometrics* 2018; 74: 801-2.
18. Johnson CE, Weerasuria MP, Keating JL. Effect of face-to-face verbal feedback compared with no or alternative feedback on the objective workplace task performance of health professionals: a systematic review and meta-analysis. *BMJ Open* 2020; 10: e030672.
19. van Aert RC, Wicherts JM, van Assen MA. Conducting meta-analyses based on p values: reservations and recommendations for applying p-uniform and p-curve. *Perspect Psychol Sci* 2016; 11: 713-29.
20. Page MJ, Higgins JP. Rethinking the assessment of risk of bias due to selective reporting: a cross-sectional study. *Syst Rev* 2016; 5: 108.
21. Jackson JL, Balk EM, Hyun N, Kuriyama A. Approaches to assessing and adjusting for

- selective outcome reporting in meta-analysis. *J Gen Intern Med* 2022; 37: 1247-53.
22. Jüni P, Reichenbach S, Egger M. COX 2 inhibitors, traditional NSAIDs, and the heart. *BMJ* 2005; 330: 1342-3.
 23. Garner SE, Fidan DD, Frankish RR, Judd MG, Towheed TE, Wells G, et al. Rofecoxib for rheumatoid arthritis. *Cochrane Database Syst Rev* 2005; 2005: CD003685.
 24. Hazlewood GS. Risk/benefit trade-offs in rheumatology: rofecoxib revisited in the era of JAK inhibitors. *Rheumatology (Oxford)* 2023; 62: 3773-5.
 25. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 2000; 320: 1574-7.
 26. Kang H. Statistical considerations in meta-analysis. *Hanyang Med Rev* 2015; 35: 23-32.
 27. Simmonds M. Quantifying the risk of error when interpreting funnel plots. *Syst Rev* 2015; 4: 24.
 28. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001; 323: 101-5.
 29. Rafla C, Zitek T, Shalaby M. A systematic review of the respiratory effects of ultrasound-guided phrenic nerve block. *Anesth Pain Med (Seoul)* 2025; 20: 371-83.
 30. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315: 629-34.
 31. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; 50: 1088-101.
 32. Simonsohn U, Nelson LD, Simmons JP. p-curve and effect size: correcting for publication bias using only significant results. *Perspect Psychol Sci* 2014; 9: 666-81.
 33. Bergman H, Henschke N, Arevalo-Rodriguez I, Buckley BS, Crosbie EJ, Davies JC, et al. Human papillomavirus (HPV) vaccination for the prevention of cervical cancer and other

HPV-related diseases: a network meta-analysis. *Cochrane Database Syst Rev* 2025; 11: CD015364.

34. Ximenes GF, Costa ÁL, Leite LL, Costa LL, Ribeiro MO, Baima Colares PG, et al. Are artificial intelligence models reliable for clinical application in pediatric fracture detection on radiographs? A systematic review and meta-analysis. *Clin Orthop Relat Res* 2026; 484: 371–85.
35. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Res Synth Methods* 2014; 5: 60-78.
36. Alinaghi N, Reed WR. Meta-analysis and publication bias: how well does the FAT-PET-PEESE procedure work? *Res Synth Methods* 2018; 9: 285-311.
37. Larose DT, Dey DK. Modeling publication bias using weighted distributions in a Bayesian framework. *Comput Stat Data Anal* 1998; 26: 279-302.
38. Maier M, Bartoš F, Wagenmakers EJ. Robust Bayesian meta-analysis: addressing publication bias with model-averaging. *Psychol Methods* 2023; 28: 107-22.
39. Jung J, Aloe AM. Bayesian workflow for bias-adjustment model in meta-analysis. *Res Synth Methods* 2026; 17: 293–313.
40. Cansian JM, Bracht VS, Biolo LV, Schmidt AP. Comparative effectiveness of bupivacaine and lidocaine-bupivacaine mixtures in brachial plexus block: a systematic review and meta-analysis. *Anesth Pain Med (Seoul)* 2025; 20: 252-65.
41. Spitschan M, Schmidt MH, Blume C. Principles of open, transparent and reproducible science in author guidelines of sleep research and chronobiology journals. *Wellcome Open Res* 2020; 5: 172.
42. Atakpo P, Vassar M. Publication bias in dermatology systematic reviews and meta-analyses. *J Dermatol Sci* 2016; 82: 69-74.

43. Mathur MB. Sensitivity analysis for the interactive effects of internal bias and publication bias in meta-analyses. *Res Synth Methods* 2024; 15: 21-43.
44. Mahendru K, Kumar A, Pandey K, Sarma R. Comparison of the effects of remimazolam and inhalational anesthesia on postoperative recovery in patients undergoing general anesthesia: a systematic review and meta-analysis of randomized controlled trials. *Anesth Pain Med (Seoul)* 2025; 20: 393-405.
45. Malmivaara A. Methodological considerations of the GRADE method. *Ann Med* 2015; 47: 1-5.
46. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeftang M, Murad MH, et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. *J Clin Epidemiol* 2020; 122: 142-52.
47. Llanaj E, Muka T. Misleading meta-analyses during COVID-19 pandemic: examples of methodological biases in evidence synthesis. *J Clin Med* 2022; 11: 4084.
48. Kang H, Oh HC. Current concerns on journal article with preprint: Anesthesia and Pain Medicine perspectives. *Anesth Pain Med (Seoul)* 2023; 18: 97-103.
49. Munn Z, Pollock D, Barker TH, Stone J, Stern C, Aromataris E, et al. The Pandora's box of evidence synthesis and the case for a living evidence synthesis taxonomy. *BMJ Evid Based Med* 2023; 28: 148-50.
50. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. Bias modelling in evidence synthesis. *J R Stat Soc Ser A Stat Soc* 2009; 172: 21-47.
51. Xing X, Lin L, Murad MH, Tong J. Leveraging AI for meta-analysis: evaluating LLMs in detecting publication bias for next-generation evidence synthesis. *Cochrane Evid Synth Methods* 2025; 3: e70047.

Epub ahead of print

Figure legends

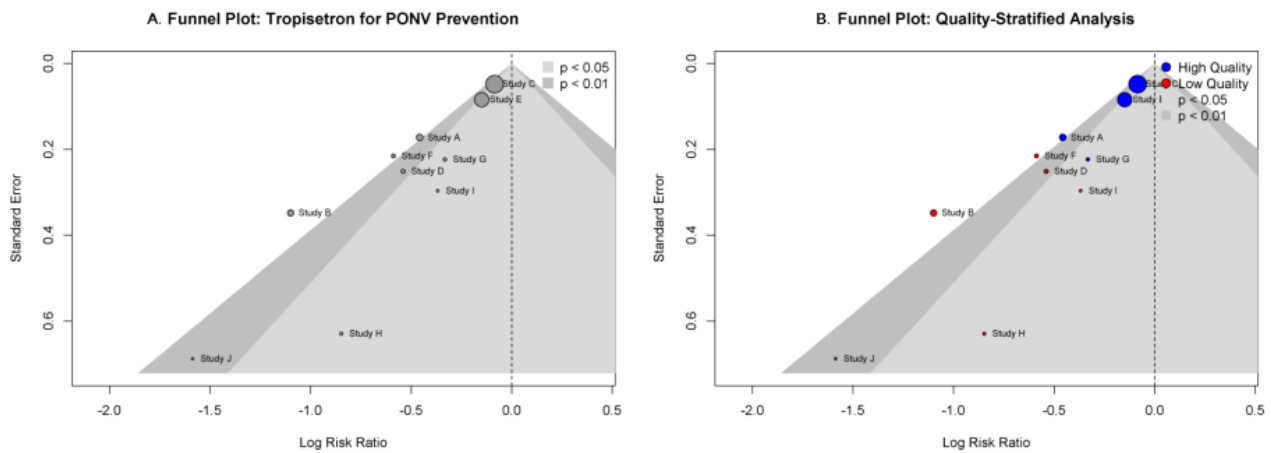


Fig. 1. Funnel plots of trials evaluating tropisetron for prevention of postoperative nausea and vomiting. (A) Conventional funnel plot. (B) Quality-stratified funnel plot. Each circle represents an individual study, plotted by log risk ratio and standard error. Shaded contours indicate regions of statistical significance ($P < 0.05$ and $P < 0.01$). Blue circles indicate higher-quality studies, whereas red circles indicate lower-quality studies.

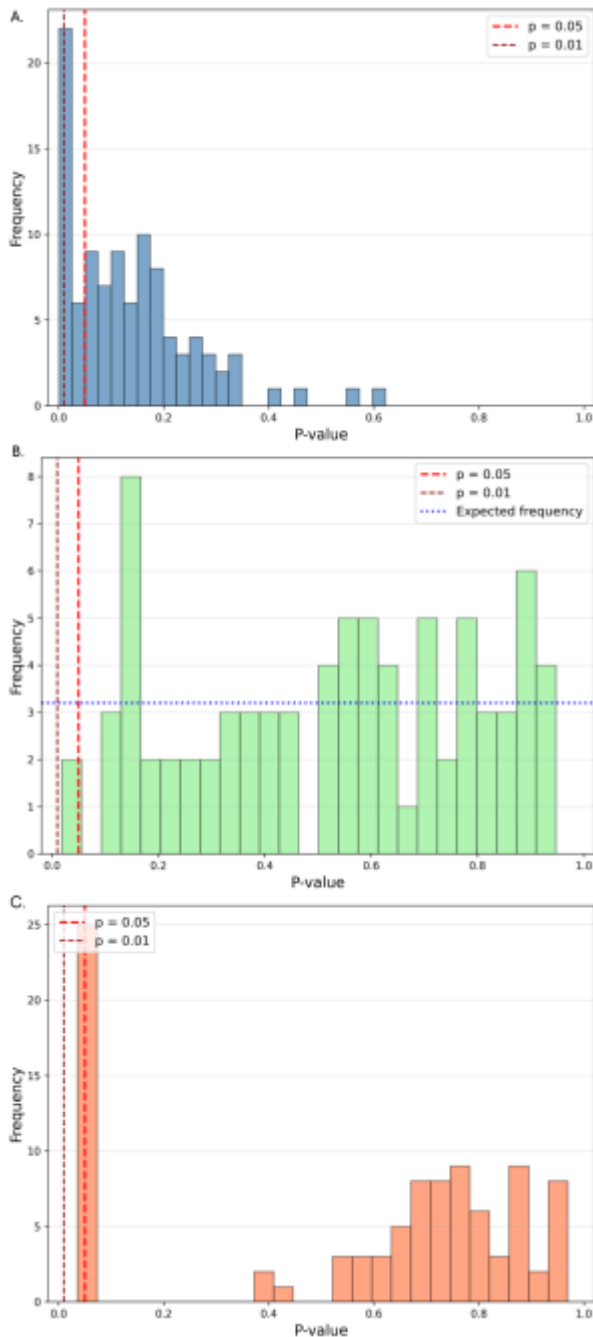


Fig. 2. Distributions of P-values under different selective reporting scenarios. (A) Distribution showing an excess of small P-values, consistent with a right-skewed pattern often interpreted as evidential value under selective reporting. (B) Distribution compared with an expected uniform frequency, illustrating how deviations from uniformity may inform assessment of selective significance filtering. (C) Distribution characterized by clustering just below conventional

significance thresholds, consistent with selective reporting or analytical flexibility rather than a coherent underlying effect. Each panel displays the frequency distribution of P-values from a set of studies. Vertical dashed lines indicate conventional significance thresholds ($P = 0.05$ and $P = 0.01$).

Epub ahead of print

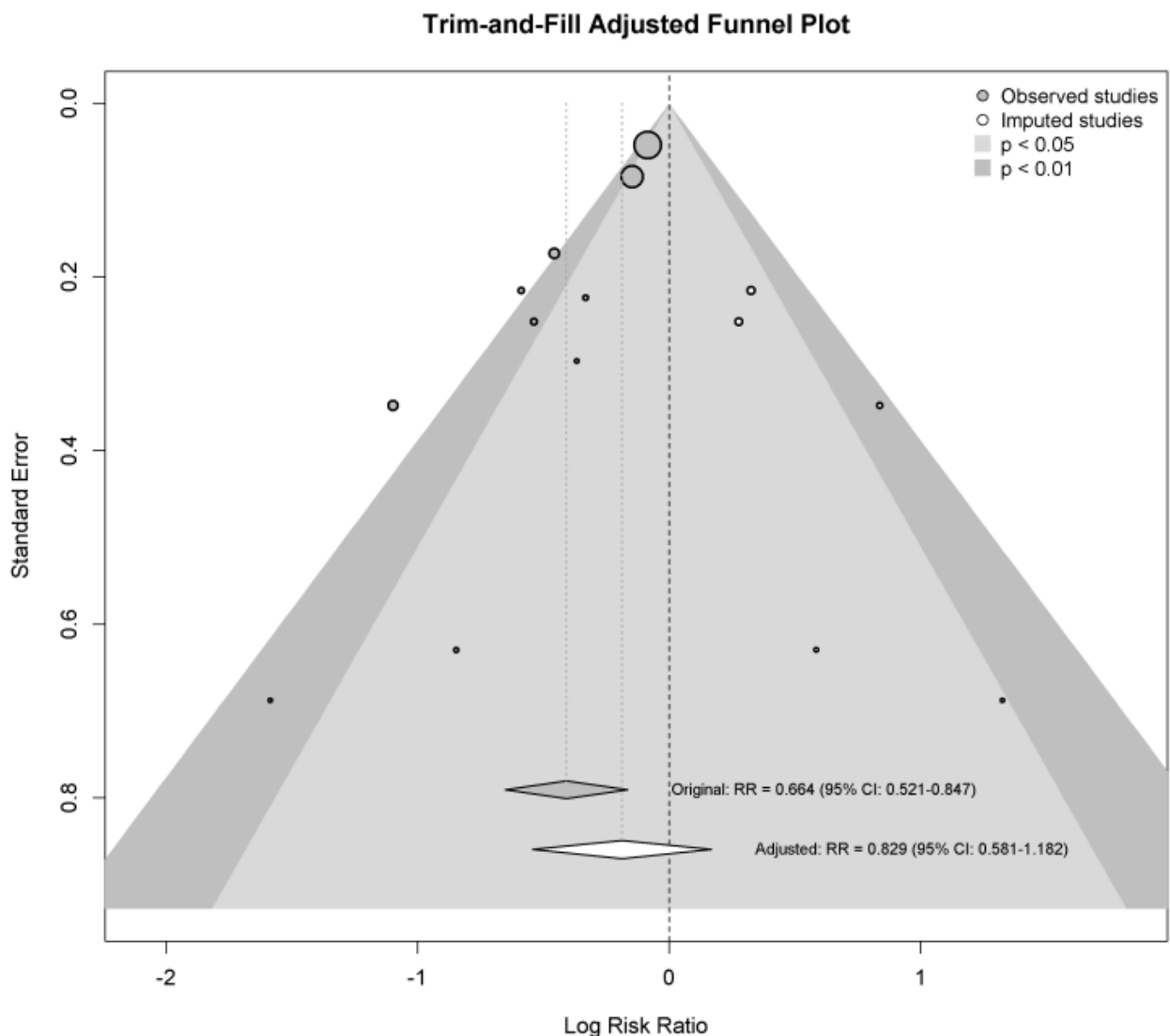


Fig. 3. Trim-and-fill-adjusted funnel. Each circle represents an individual study, plotted by log risk ratio and standard error. Shaded contours indicate regions of statistical significance ($P < 0.05$ and $P < 0.01$). Open circles denote observed studies, and filled circles denote imputed studies. Diamonds indicate the pooled effect estimates before (gray) and after (white) trim-and-fill adjustment, with corresponding 95% CI.

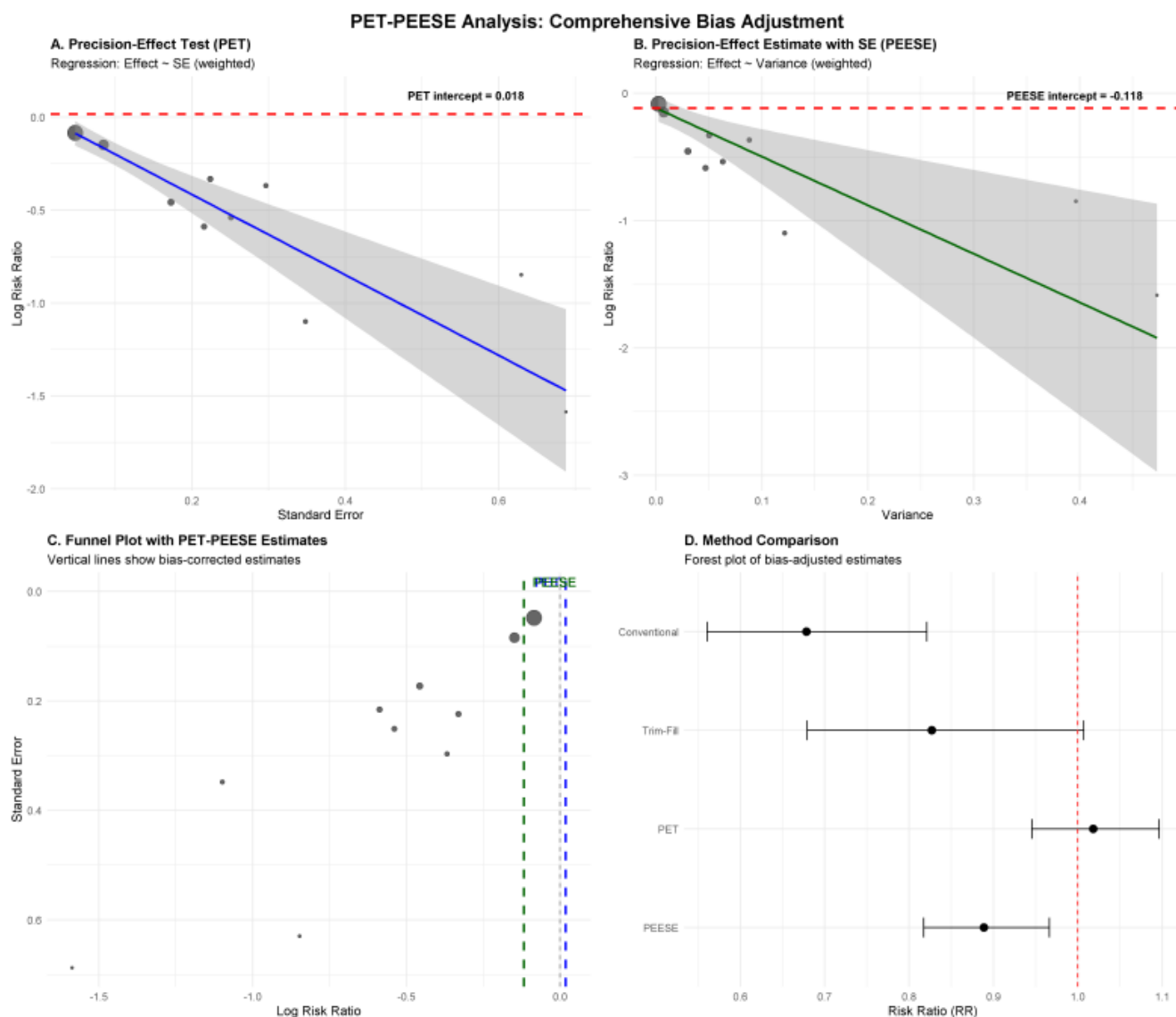


Fig. 4. PET–PEESE regression adjustments for publication bias. (A) Precision-effect test (PET), showing regression of log risk ratio on standard error. (B) Precision-effect estimate with standard errors (PEESE), showing regression of log risk ratio on variance. (C) Comparison of pooled effect estimates from the conventional random-effects meta-analysis, PET, and PEESE, displayed as risk ratios with 95% CI. (D) Scatter plot of treatment effect by study precision (1/standard error), illustrating the association between study size and estimated effect. In panels A and B, each circle represents an individual study, with point size proportional to study precision. Dashed horizontal lines indicate the conventional random-effects estimate and the PET- or PEESE-adjusted estimates. Blue

dashed lines denote the conventional random-effects estimate, red dashed lines denote the PET-adjusted estimate, and green dashed lines denote the PEESE-adjusted estimate.

Epub ahead of print

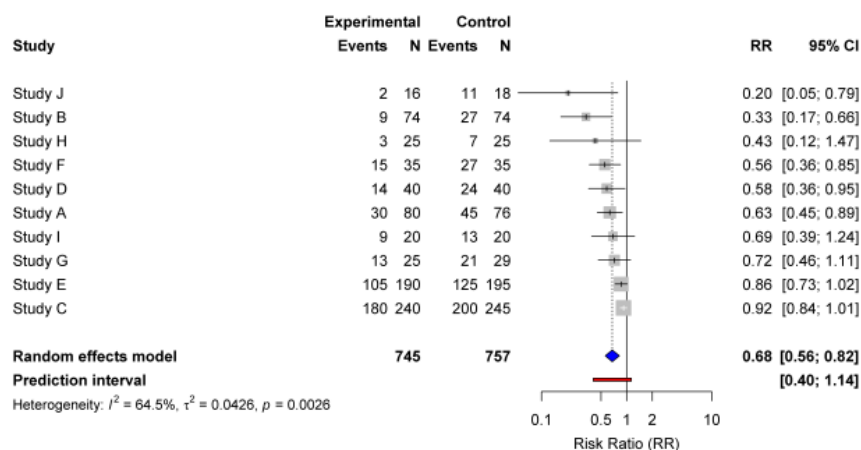


Fig. 5. Forest plot for preventive effect of postoperative nausea and vomiting. Each trial is shown as a filled square proportional to sample size, with the 95% CI depicted as a horizontal line. The blue diamond indicates the pooled effect and 95% CI and the red bar indicates 95% prediction interval. N number, RR risk ratio.

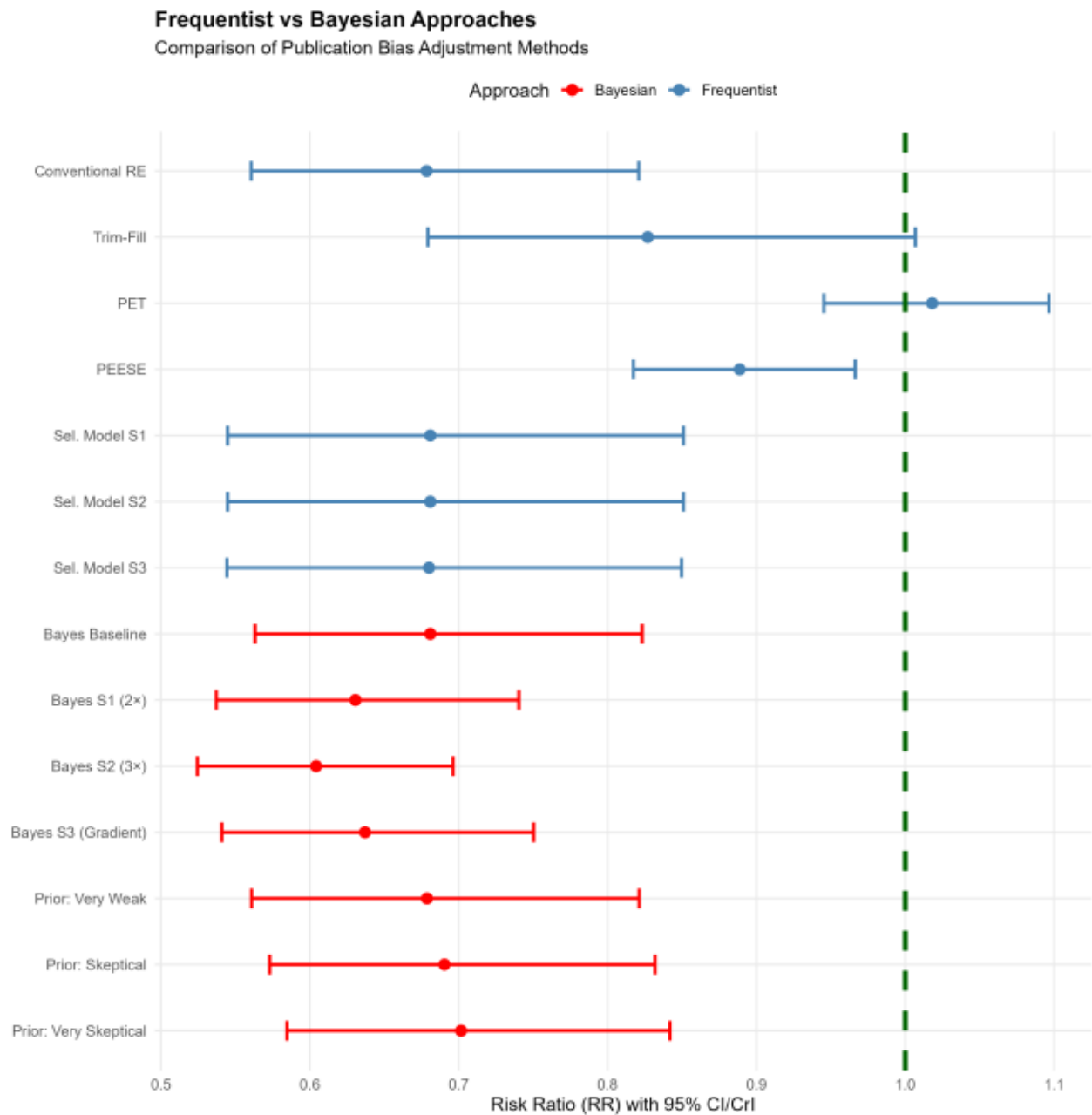


Fig. 6. Comparison of frequentist and Bayesian approaches to publication bias adjustment.

Forest plot comparing publication bias adjustment methods. Frequentist (blue) and Bayesian (red) approaches show varying estimates based on different assumptions. Green line indicates null effect (RR = 1.0). Error bars: 95% CI (frequentist) or CrI (Bayesian). Substantial divergence across methods or prior specifications indicates that the pooled estimate is sensitive to modeling assumptions and should be interpreted with caution. CrI: credible interval

Fig. 7. P-value distribution and registry-based evidence of selective reporting. (A) Right-skewed p-curve among significant studies suggests evidential value. (B) Registry comparison reveals outcome switching or multiplicity. Red dashed lines indicate $P = 0.025$ (A) and $P = 0.05$ (B) thresholds.

Epub ahead of print

```
#####
#
# SUPPLEMENTARY R CODE: Comprehensive Publication Bias Analysis
# For PONV Meta-analysis
#
# This script reproduces all analyses and figures from the manuscript
#
#####

# =====
# SETUP AND INSTALLATION
# =====

# Set your working directory (modify this path as needed)
# setwd("your/path/here")

rm(list = ls())

# 예: 해당 폴더를 작업 디렉터리로 지정
setwd("E:/Hard/연구들/Publication bias/Analysis10")
getwd()

library(meta)
library(ggplot2)
library(gridExtra)
library(grid)
library(metafor)

# =====
# 1. DATA PREPARATION
# =====

ponv_data <- data.frame(
  study_id = c("Study A", "Study B", "Study C", "Study D",
```

```

        "Study E", "Study F", "Study G",
        "Study H", "Study I", "Study J"),
event.e = c(30, 9, 180, 14, 105, 15, 13, 3, 9, 2),
n.e = c(80, 74, 240, 40, 190, 35, 25, 25, 20, 16),
event.c = c(45, 27, 200, 24, 125, 27, 21, 7, 13, 11),
n.c = c(76, 74, 245, 40, 195, 35, 29, 25, 20, 18)
)

print(ponv_data)

# =====
# 2. CONVENTIONAL RANDOM-EFFECTS META-ANALYSIS (DerSimonian-Laird 방법)
# =====

m.ponv <- metabin(
  event.e = event.e,
  n.e = n.e,
  event.c = event.c,
  n.c = n.c,
  studlab = study_id,
  data = ponv_data,
  sm = "RR",
  method = "Inverse",
  MH.exact = TRUE,
  fixed = FALSE,
  random = TRUE,
  method.tau = "DL",
  method.random.ci = "classic",
  title = "PONV"
)

print(m.ponv$tau^2)
print(m.ponv$I2)

```

```

print(exp(m.ponv$TE.random))
print(exp(m.ponv$lower.random))
print(exp(m.ponv$upper.random))

# =====
# 3. FIGURE 5: FOREST PLOT
# =====

png("Figure_5_Forest_Plot.png", width = 14, height = 10, units = "in", res = 600)
par(mar = c(5, 4, 4, 2))
forest(m.ponv,
      sortvar = TE,
      prediction = TRUE,
      print.tau2 = TRUE,
      leftcols = c("studlab", "event.e", "n.e", "event.c", "n.c"),
      leftlabs = c("Study", "Events", "N", "Events", "N"),
      rightcols = c("effect", "ci"),
      rightlabs = c("RR", "95% CI"),
      xlab = "Risk Ratio (RR)",
      smlab = "",
      weight.study = "random",
      col.diamond = "blue",
      col.predict = "red",
      spacing = 1.3,
      cex = 0.8,
      cex.lab = 0.9,
      calcwidth.hetstat = TRUE,
      calcwidth.overall = TRUE)
dev.off()

# =====

```

#Funnel Plot을 위한 준비작업

```
# =====  
ponv_data$total_n <- ponv_data$n.e + ponv_data$n.c  
sample_sizes <- ponv_data$total_n  
point_sizes <- 0.5 + 2.5 * (sample_sizes - min(sample_sizes)) / (max(sample_sizes)  
- min(sample_sizes))
```

```
# =====
```

#기본 Funnel Plot Figure 1A

```
# =====
```

```
png("Figure_1A_Funnel_Plot.png", width = 8, height = 6, units = "in", res = 600)  
par(mar = c(5, 4, 4, 2) + 0.1)
```

```
plot(m.ponv$TE, m.ponv$seTE,  
     type = "n",  
     xlim = range(m.ponv$TE) + c(-0.5, 0.5),  
     ylim = c(max(m.ponv$seTE) * 1.05, 0),  
     xlab = "Log Risk Ratio",  
     ylab = "Standard Error",  
     main = "Funnel Plot: Tropisetron for PONV Prevention")
```

```
se_seq <- seq(0, max(m.ponv$seTE) * 1.05, length.out = 100)
```

```
polygon(c(-2.576 * se_seq, rev(2.576 * se_seq)),  
        c(se_seq, rev(se_seq)),  
        col = "gray75", border = NA)
```

```
polygon(c(-1.96 * se_seq, rev(1.96 * se_seq)),  
        c(se_seq, rev(se_seq)),  
        col = "gray85", border = NA)
```

```
abline(v = 0, lty = 2)
```

```

points(m.ponv$TE, m.ponv$seTE,
      pch = 21, bg = "gray60", col = "black",
      cex = point_sizes, lwd = 1)

text(m.ponv$TE, m.ponv$seTE,
     labels = m.ponv$studlab,
     pos = 4, cex = 0.7, offset = 0.5)

legend("topright",
      legend = c("p < 0.05", "p < 0.01"),
      pch = 15, col = c("gray85", "gray75"),
      pt.cex = 1.8, bty = "n")

dev.off()
cat("Figure 1A saved!\n")

# =====
# 4B. FIGURE 1B: QUALITY-STRATIFIED FUNNEL PLOT
# =====

quality <- ifelse(ponv_data$study_id == "Study B", "Low Quality",
                 ifelse(ponv_data$total_n > 200, "High Quality",
                       ifelse(ponv_data$total_n < 100 & exp(m.ponv$TE) < 0.7, "Low
Quality", "High Quality")))
quality_colors <- ifelse(quality == "High Quality", "blue", "red")

#data.frame(Study = ponv_data$study_id, Quality = quality)

png("Figure_1B_Funnel_Quality.png", width = 8, height = 6, units = "in", res =
600)

par(mar = c(5, 4, 4, 2) + 0.1)

```

```

plot(m.ponv$TE, m.ponv$seTE,
     type = "n",
     xlim = range(m.ponv$TE) + c(-0.5, 0.5),
     ylim = c(max(m.ponv$seTE) * 1.05, 0),
     xlab = "Log Risk Ratio",
     ylab = "Standard Error",
     main = "Funnel Plot: Quality-Stratified Analysis")

se_seq <- seq(0, max(m.ponv$seTE) * 1.05, length.out = 100)

polygon(c(-2.576 * se_seq, rev(2.576 * se_seq)),
        c(se_seq, rev(se_seq)),
        col = "gray75", border = NA)

polygon(c(-1.96 * se_seq, rev(1.96 * se_seq)),
        c(se_seq, rev(se_seq)),
        col = "gray85", border = NA)

abline(v = 0, lty = 2)

points(m.ponv$TE, m.ponv$seTE,
       pch = 21, bg = quality_colors, col = "black",
       cex = point_sizes, lwd = 1)

text(m.ponv$TE, m.ponv$seTE,
     labels = m.ponv$studlab,
     pos = 4, cex = 0.7, offset = 0.5)

legend("topright",
      legend = c("High Quality", "Low Quality", "p < 0.05", "p < 0.01"),
      pch = c(21, 21, 15, 15),
      pt.bg = c("blue", "red", "gray85", "gray75"),
      col = c("black", "black", "gray85", "gray75"),

```

```

pt.cex = 1.5, bty = "n")

dev.off()

cat("Figure 1B saved!\n")

# =====
# 5. PUBLICATION BIAS DETECTION
# =====

egger_test <- metabias(m.ponv, method.bias = "linreg", k.min = 3)
cat("Egger's Test:\n")
print(egger_test)

begg_test <- metabias(m.ponv, method.bias = "rank", k.min = 3)
cat("\nBegg's Test:\n")
print(begg_test)

# =====
# 6. FIGURE 3: TRIM-AND-FILL (통계적으로 올바른 크기)
# =====

m.trimfill <- trimfill(m.ponv)
print(summary(m.trimfill))

png("Figure_3_Trim_Fill.png", width = 10, height = 9, units = "in", res = 600)
par(mar = c(5, 4, 4, 2))

plot(m.trimfill$TE, m.trimfill$seTE,
     type = "n",
     xlim = range(m.trimfill$TE) + c(-0.5, 0.5),
     ylim = c(max(m.trimfill$seTE) * 1.35, 0),
     xlab = "Log Risk Ratio",
     ylab = "Standard Error",

```

```

main = "Trim-and-Fill Adjusted Funnel Plot",
cex.main = 1.3,
cex.lab = 1.1,
cex.axis = 1.1)

# Contours
se_seq <- seq(0, max(m.trimfill$seTE) * 1.35, length.out = 100)

polygon(c(-2.576 * se_seq, rev(2.576 * se_seq)),
        c(se_seq, rev(se_seq)),
        col = "gray75", border = NA)

polygon(c(-1.96 * se_seq, rev(1.96 * se_seq)),
        c(se_seq, rev(se_seq)),
        col = "gray85", border = NA)

abline(v = 0, lty = 2)

# Calculate sizes
n_observed <- length(m.ponv$TE)
n_total <- length(m.trimfill$TE)
n_imputed <- n_total - n_observed

all_sizes <- 0.5 + 2.5 * (1/m.trimfill$seTE - min(1/m.trimfill$seTE)) /
  (max(1/m.trimfill$seTE) - min(1/m.trimfill$seTE))

# Observed studies (gray)
points(m.trimfill$TE[1:n_observed], m.trimfill$seTE[1:n_observed],
       pch = 21, bg = "gray", col = "black",
       cex = point_sizes, lwd = 2)

# Imputed studies (white)
if(n_imputed > 0) {

```

```

imputed_indices <- (n_observed + 1):n_total
points(m.trimfill$TE[imputed_indices], m.trimfill$seTE[imputed_indices],
      pch = 21, bg = "white", col = "black",
      cex = all_sizes[imputed_indices], lwd = 2)
}

# Diamonds
usr <- par("usr")
y_max <- max(m.trimfill$seTE, na.rm = TRUE)
diamond_y1 <- y_max * 1.15
diamond_y2 <- y_max * 1.25
diamond_height <- y_max * 0.015

polygon(x = c(m.ponv$TE.random, m.ponv$upper.random,
             m.ponv$TE.random, m.ponv$lower.random),
       y = c(diamond_y1 - diamond_height, diamond_y1,
             diamond_y1 + diamond_height, diamond_y1),
       col = "gray", border = "black", lwd = 1.5)

polygon(x = c(m.trimfill$TE.random, m.trimfill$upper.random,
             m.trimfill$TE.random, m.trimfill$lower.random),
       y = c(diamond_y2 - diamond_height, diamond_y2,
             diamond_y2 + diamond_height, diamond_y2),
       col = "white", border = "black", lwd = 1.5)

segments(m.ponv$TE.random, 0, m.ponv$TE.random, diamond_y1 - diamond_height,
        lty = 3, col = "darkgray", lwd = 1.5)
segments(m.trimfill$TE.random, 0, m.trimfill$TE.random, diamond_y2 -
diamond_height,
        lty = 3, col = "darkgray", lwd = 1.5)

text(m.ponv$upper.random + (usr[2] - usr[1]) * 0.03, diamond_y1,
     paste0("Original: RR = ", round(exp(m.ponv$TE.random), 3),
           " (95% CI: ", round(exp(m.ponv$lower.random), 3), "-",

```

```

        round(exp(m.ponv$upper.random), 3), ")),
pos = 4, cex = 0.8)

text(m.trimfill$upper.random + (usr[2] - usr[1]) * 0.03, diamond_y2,
     paste0("Adjusted: RR = ", round(exp(m.trimfill$TE.random), 3),
           " (95% CI: ", round(exp(m.trimfill$lower.random), 3), "-",
           round(exp(m.trimfill$upper.random), 3), "))),
pos = 4, cex = 0.8)

legend("topright",
      legend = c("Observed studies", "Imputed studies", "p < 0.05", "p < 0.01"),
      pch = c(21, 21, 15, 15),
      pt.bg = c("gray", "white", "gray85", "gray75"),
      col = c("black", "black", "gray85", "gray75"),
      pt.cex = c(1.2, 1.2, 1.8, 1.8),
      bty = "n")

dev.off()
cat("Figure 3 saved!\n")

# =====
# 8. SELECTION MODEL SCENARIOS
# =====

library(metafor)

ponv_es <- escalc(measure = "RR",
                 ai = event.e, n1i = n.e,
                 ci = event.c, n2i = n.c,
                 data = ponv_data)

```

```

z_values <- ponv_es$yi / sqrt(ponv_es$vi)
p_values <- 2 * (1 - pnorm(abs(z_values)))
rma_model <- rma(yi = yi, vi = vi, data = ponv_es, method = "DL")

# Scenario 1
cat("Scenario 1: Moderate selection (p<0.05: 2×)\n")
sel_scenario1 <- selmodel(rma_model,
                          type = "stepfun",
                          steps = 0.05,
                          delta = c(2, 1))
print(sel_scenario1)

cat(" RR =", round(exp(sel_scenario1$beta), 3), "\n")
cat(" 95% CI: [", round(exp(sel_scenario1$ci.lb), 3), "-",
    round(exp(sel_scenario1$ci.ub), 3), "]\n\n")

#Scenario 2:
cat("Scenario 2: Strong selection (p<0.05: 3×)\n")
sel_scenario2 <- selmodel(rma_model,
                          type = "stepfun",
                          steps = 0.05,
                          delta = c(3, 1))
print(sel_scenario2)

cat(" RR =", round(exp(sel_scenario2$beta), 3), "\n")
cat(" 95% CI: [", round(exp(sel_scenario2$ci.lb), 3), "-",
    round(exp(sel_scenario2$ci.ub), 3), "]\n\n")

#Scenario 3:
cat("Scenario 3: Custom gradient\n")
sel_scenario3 <- selmodel(rma_model,
                          type = "stepfun",
                          steps = c(0.01, 0.05),

```

```

        delta = c(2, 1.5, 1))

print(sel_scenario3)

cat("  RR =", round(exp(sel_scenario3$beta), 3), "\n")
cat("  95% CI: [", round(exp(sel_scenario3$ci.lb), 3), "-",
    round(exp(sel_scenario3$ci.ub), 3), "]\n\n")

#####

cat("=== 각 연구의 P-value ===\n\n")

p_table <- data.frame(
  Study = ponv_data$study_id,
  P_value = round(p_values, 4),
  Category = ifelse(p_values < 0.01, "p < 0.01 (2배)",
    ifelse(p_values < 0.05, "0.01 ≤ p < 0.05 (1.5배)",
      "p ≥ 0.05 (1배)"))
)

print(p_table)

cat("\n분포 요약:\n")

cat("  p < 0.01:", sum(p_values < 0.01), "개\n")

cat("  0.01 ≤ p < 0.05:", sum(p_values >= 0.01 & p_values < 0.05), "개\n")

cat("  p ≥ 0.05:", sum(p_values >= 0.05), "개\n")

# =====
# 9. PET-PEESE ANALYSIS
# =====

#STEP 1: PET (Precision-Effect Test)

```

```

cat("=== PET-PEESE Analysis ===\n\n")

pet_model <- lm(ponv_es$yi ~ sqrt(ponv_es$vi), weights = 1/ponv_es$vi)

cat("PET Model:\n")
print(summary(pet_model))

log_rr_pet <- coef(pet_model)[1]
se_pet <- summary(pet_model)$coefficients[1, 2]
rr_pet <- exp(log_rr_pet)

cat("\nPET Estimate:\n")
cat("  RR =", round(rr_pet, 3), "\n")
cat("  95% CI: [", round(exp(log_rr_pet - 1.96*se_pet), 3), "-",
    round(exp(log_rr_pet + 1.96*se_pet), 3), "]\n\n")

#STEP 2: PEESE (Precision-Effect Estimate with Standard Error)

peese_model <- lm(ponv_es$yi ~ ponv_es$vi, weights = 1/ponv_es$vi)

cat("PEESE Model:\n")
print(summary(peese_model))

log_rr_peese <- coef(peese_model)[1]
se_peese <- summary(peese_model)$coefficients[1, 2]
rr_peese <- exp(log_rr_peese)

cat("\nPEESE Estimate:\n")
cat("  RR =", round(rr_peese, 3), "\n")
cat("  95% CI: [", round(exp(log_rr_peese - 1.96*se_peese), 3), "-",
    round(exp(log_rr_peese + 1.96*se_peese), 3), "]\n\n")

```

```

# =====
# FIGURE: 4-PANEL PET-PEESE VISUALIZATION
# =====

pet_df <- data.frame(
  se = sqrt(ponv_es$vi),
  log_rr = ponv_es$yi,
  var = ponv_es$vi,
  weight = 1/ponv_es$vi
)

# Panel A: PET regression
pet_panel_a <- ggplot(pet_df, aes(x = se, y = log_rr)) +
  geom_point(aes(size = weight), alpha = 0.6) +
  geom_smooth(method = "lm", aes(weight = weight),
    formula = y ~ x, se = TRUE, color = "blue") +
  geom_hline(yintercept = log_rr_pet, linetype = "dashed", color = "red",
    linewidth = 1) +
  annotate("text", x = max(pet_df$se) * 0.7, y = log_rr_pet + 0.1,
    label = paste0("PET intercept = ", round(log_rr_pet, 3)),
    hjust = 0, size = 3.5, fontface = "bold") +
  labs(title = "A. Precision-Effect Test (PET)",
    subtitle = "Regression: Effect ~ SE (weighted)",
    x = "Standard Error", y = "Log Risk Ratio") +
  theme_minimal() +
  theme(legend.position = "none",
    plot.title = element_text(face = "bold", size = 12))

# Panel B: PEESE regression
pet_panel_b <- ggplot(pet_df, aes(x = var, y = log_rr)) +
  geom_point(aes(size = weight), alpha = 0.6) +
  geom_smooth(method = "lm", aes(weight = weight),
    formula = y ~ x, se = TRUE, color = "darkgreen") +

```

```

  geom_hline(yintercept = log_rr_peese, linetype = "dashed", color = "red",
linewidth = 1) +

  annotate("text", x = max(pet_df$var) * 0.7, y = log_rr_peese + 0.1,
    label = paste0("PEESE intercept = ", round(log_rr_peese, 3)),
    hjust = 0, size = 3.5, fontface = "bold") +

  labs(title = "B. Precision-Effect Estimate with SE (PEESE)",
    subtitle = "Regression: Effect ~ Variance (weighted)",
    x = "Variance", y = "Log Risk Ratio") +

  theme_minimal() +

  theme(legend.position = "none",
    plot.title = element_text(face = "bold", size = 12))

# Panel C: Funnel plot with regression lines
pet_panel_c <- ggplot(pet_df, aes(x = log_rr, y = se)) +
  geom_point(aes(size = weight), alpha = 0.6) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
  geom_vline(xintercept = log_rr_pet, linetype = "dashed", color = "blue",
linewidth = 1) +
  geom_vline(xintercept = log_rr_peese, linetype = "dashed", color = "darkgreen",
linewidth = 1) +

  scale_y_reverse() +

  annotate("text", x = log_rr_pet, y = 0, label = "PET",
    vjust = -0.5, hjust = 1.2, color = "blue", fontface = "bold") +

  annotate("text", x = log_rr_peese, y = 0, label = "PEESE",
    vjust = -0.5, hjust = -0.2, color = "darkgreen", fontface = "bold") +

  labs(title = "C. Funnel Plot with PET-PEESE Estimates",
    subtitle = "Vertical lines show bias-corrected estimates",
    x = "Log Risk Ratio", y = "Standard Error") +

  theme_minimal() +

  theme(legend.position = "none",
    plot.title = element_text(face = "bold", size = 12))

# Panel D: Comparison of estimates
comparison_data <- data.frame(

```

```

Method = c("Conventional", "Trim-Fill", "PET", "PEESE"),
RR = c(exp(m.ponv$TE.random), exp(m.trimfill$TE.random), rr_pet, rr_peese),
Lower = c(exp(m.ponv$lower.random), exp(m.trimfill$lower.random),
           exp(log_rr_pet - 1.96*se_pet), exp(log_rr_peese - 1.96*se_peese)),
Upper = c(exp(m.ponv$upper.random), exp(m.trimfill$upper.random),
           exp(log_rr_pet + 1.96*se_pet), exp(log_rr_peese + 1.96*se_peese))
)
comparison_data$Method <- factor(comparison_data$Method,
                                levels = rev(c("Conventional", "Trim-Fill", "PET",
"PEESE"))))

pet_panel_d <- ggplot(comparison_data, aes(x = RR, y = Method)) +
  geom_point(size = 3) +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.2) +
  geom_vline(xintercept = 1, linetype = "dashed", color = "red") +
  labs(title = "D. Method Comparison",
       subtitle = "Forest plot of bias-adjusted estimates",
       x = "Risk Ratio (RR)", y = "") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 12))

# Save combined figure
png("figure4_pet_peese_comprehensive.png", width = 14, height = 12, units = "in",
    res = 300)
grid.arrange(pet_panel_a, pet_panel_b, pet_panel_c, pet_panel_d, ncol = 2,
             top = textGrob("PET-PEESE Analysis: Comprehensive Bias Adjustment",
                           gp = gpar(fontsize = 16, fontface = "bold")))
dev.off()
cat("PET-PEESE figure saved: figure4_pet_peese_comprehensive.png\n\n")

#####
#####
#Bayesian Analysis
#####

```

```
#####
```

```
cat("=== Bayesian Analysis ===\n\n")
```

```
ponv_es <- escalc(measure = "RR",  
                 ai = event.e, n1i = n.e,  
                 ci = event.c, n2i = n.c,  
                 data = ponv_data)
```

```
ponv_rma <- rma(yi, vi, data = ponv_es, method = "DL")
```

```
tau2 <- ponv_rma$tau2  
cat("Tau2 =", round(tau2, 4), "\n")
```

```
# Calculate log RR and SE
```

```
log_rr <- log((ponv_data$event.e/ponv_data$n.e) /  
              (ponv_data$event.c/ponv_data$n.c))  
se_rr <- sqrt(1/ponv_data$event.e + 1/ponv_data$event.c -  
              1/ponv_data$n.e - 1/ponv_data$n.c)
```

```
# tau2 정의
```

```
tau2 <- ponv_rma$tau2
```

```
# pooled_re 정의
```

```
weights_re <- 1 / (se_rr^2 + tau2)  
pooled_re <- sum(weights_re * log_rr) / sum(weights_re)
```

```
cat("Pooled log RR =", round(pooled_re, 3), "\n")
```

```
cat("Tau2 =", round(tau2, 4), "\n")
```

```
pooled_re <- sum(weights_re * log_rr) / sum(weights_re)
```

```
# Fixed-effect for heterogeneity
```

```
weights_fe <- 1 / (se_rr^2)  
pooled_fe <- sum(weights_fe * log_rr) / sum(weights_fe)  
tau2 <- ponv_rma$tau2
```

```

cat("Tau² =", round(tau2, 4), "\n")
cat("Pooled log RR =", round(pooled_re, 3), "\n")

# Baseline parameters
data_precision_re <- sum(weights_re)
prior_mean <- 0 # Weakly informative prior centered at null
prior_precision <- 1 # Weak prior

# Posterior for baseline (no selection)
posterior_precision_base <- prior_precision + data_precision_re
posterior_mean_base <- (prior_precision * prior_mean +
                        data_precision_re * pooled_re) / posterior_precision_base
posterior_se_base <- sqrt(1 / posterior_precision_base)

# Calculate tau2 (DerSimonian-Laird)
Q <- sum(weights_fe * (log_rr - pooled_fe)^2)
df_Q <- length(log_rr) - 1
C <- sum(weights_fe) - sum(weights_fe^2) / sum(weights_fe)
tau2 <- max(0, (Q - df_Q) / C)

# Random-effects weights
weights_re <- 1 / (se_rr^2 + tau2)
pooled_re <- sum(weights_re * log_rr) / sum(weights_re)

cat("준비 완료!\n")

rr_bayes_base <- exp(posterior_mean_base)
lower_bayes_base <- exp(posterior_mean_base - 1.96*posterior_se_base)
upper_bayes_base <- exp(posterior_mean_base + 1.96*posterior_se_base)

cat("Bayesian Baseline (weakly informative prior):\n")
cat(" RR =", round(rr_bayes_base, 3), "\n")
cat(" 95% CrI: [", round(lower_bayes_base, 3), "-",
    round(upper_bayes_base, 3), "]\n\n")

```

```

#Scenario 1: Moderate selection (p<0.05: 2×)
cat("Bayesian Scenario 1: Moderate selection (p<0.05: 2×)\n")

# P-values
z_values <- log_rr / se_rr
p_values <- 2 * (1 - pnorm(abs(z_values)))

# Selection weights
weights_s1 <- ifelse(p_values < 0.05, 2, 1) * weights_re

# Bayesian update
data_precision_s1 <- sum(weights_s1)
pooled_s1 <- sum(weights_s1 * log_rr) / sum(weights_s1)

posterior_precision_s1 <- prior_precision + data_precision_s1
posterior_mean_s1 <- (prior_precision * prior_mean +
                      data_precision_s1 * pooled_s1) / posterior_precision_s1
posterior_se_s1 <- sqrt(1 / posterior_precision_s1)

rr_bayes_s1 <- exp(posterior_mean_s1)
lower_bayes_s1 <- exp(posterior_mean_s1 - 1.96*posterior_se_s1)
upper_bayes_s1 <- exp(posterior_mean_s1 + 1.96*posterior_se_s1)

cat(" RR =", round(rr_bayes_s1, 3), "\n")
cat(" 95% CrI: [", round(lower_bayes_s1, 3), "-",
    round(upper_bayes_s1, 3), "]\n\n")

#Scenario 2: Strong selection (p<0.05: 3×)
cat("Bayesian Scenario 2: Strong selection (p<0.05: 3×)\n")

# Selection weights
weights_s2 <- ifelse(p_values < 0.05, 3, 1) * weights_re

```

```

# Bayesian update
data_precision_s2 <- sum(weights_s2)
pooled_s2 <- sum(weights_s2 * log_rr) / sum(weights_s2)

posterior_precision_s2 <- prior_precision + data_precision_s2
posterior_mean_s2 <- (prior_precision * prior_mean +
                      data_precision_s2 * pooled_s2) / posterior_precision_s2
posterior_se_s2 <- sqrt(1 / posterior_precision_s2)

rr_bayes_s2 <- exp(posterior_mean_s2)
lower_bayes_s2 <- exp(posterior_mean_s2 - 1.96*posterior_se_s2)
upper_bayes_s2 <- exp(posterior_mean_s2 + 1.96*posterior_se_s2)

cat(" RR =", round(rr_bayes_s2, 3), "\n")
cat(" 95% CrI: [", round(lower_bayes_s2, 3), "-",
    round(upper_bayes_s2, 3), "]\n\n")

#Scenario 3: Custom gradient (p<0.01: 2×, 0.01≤p<0.05: 1.5×)
cat("Bayesian Scenario 3: Custom gradient\n")

# Selection weights
weights_s3 <- ifelse(p_values < 0.01, 2,
                    ifelse(p_values < 0.05, 1.5, 1)) * weights_re

# Bayesian update
data_precision_s3 <- sum(weights_s3)
pooled_s3 <- sum(weights_s3 * log_rr) / sum(weights_s3)

posterior_precision_s3 <- prior_precision + data_precision_s3
posterior_mean_s3 <- (prior_precision * prior_mean +
                      data_precision_s3 * pooled_s3) / posterior_precision_s3
posterior_se_s3 <- sqrt(1 / posterior_precision_s3)

```

```

rr_bayes_s3 <- exp(posterior_mean_s3)
lower_bayes_s3 <- exp(posterior_mean_s3 - 1.96*posterior_se_s3)
upper_bayes_s3 <- exp(posterior_mean_s3 + 1.96*posterior_se_s3)

cat(" RR =", round(rr_bayes_s3, 3), "\n")
cat(" 95% CrI: [", round(lower_bayes_s3, 3), "-",
    round(upper_bayes_s3, 3), "]\n\n")

#결과요약
cat("=== Bayesian Results Summary ===\n\n")

bayes_results <- data.frame(
  Scenario = c("Baseline", "Scenario 1 (p<0.05: 2×)",
    "Scenario 2 (p<0.05: 3×)", "Scenario 3 (Gradient)"),
  RR = c(rr_bayes_base, rr_bayes_s1, rr_bayes_s2, rr_bayes_s3),
  Lower = c(lower_bayes_base, lower_bayes_s1, lower_bayes_s2, lower_bayes_s3),
  Upper = c(upper_bayes_base, upper_bayes_s1, upper_bayes_s2, upper_bayes_s3)
)

print(bayes_results)

cat("\n")

#####
#####
#PRIOR sensitivity analysis
#####
#####
#Prior 1: Very weak prior (거의 flat)
cat("=== Prior Sensitivity Analysis ===\n\n")

```

```

# Prior 1: Very weak (거의 flat)

cat("Prior 1: Very weak prior (precision = 0.1)\n")
prior_precision_1 <- 0.1

posterior_precision_1 <- prior_precision_1 + data_precision_re
posterior_mean_1 <- (prior_precision_1 * prior_mean +
                     data_precision_re * pooled_re) / posterior_precision_1
posterior_se_1 <- sqrt(1 / posterior_precision_1)

rr_prior1 <- exp(posterior_mean_1)
lower_prior1 <- exp(posterior_mean_1 - 1.96*posterior_se_1)
upper_prior1 <- exp(posterior_mean_1 + 1.96*posterior_se_1)

cat(" RR =", round(rr_prior1, 3), "\n")
cat(" 95% CrI: [", round(lower_prior1, 3), "-",
    round(upper_prior1, 3), "]\n\n")

#Prior 2: Moderate prior (기본값)

cat("Prior 2: Moderate prior (precision = 1) - Baseline\n")
cat(" RR =", round(rr_bayes_base, 3), "\n")
cat(" 95% CrI: [", round(lower_bayes_base, 3), "-",
    round(upper_bayes_base, 3), "]\n\n")

#Prior 3: Skeptical prior (효과 없다고 가정)

cat("Prior 3: Skeptical prior (precision = 5, mean = 0)\n")
prior_precision_3 <- 5

posterior_precision_3 <- prior_precision_3 + data_precision_re
posterior_mean_3 <- (prior_precision_3 * 0 +
                     data_precision_re * pooled_re) / posterior_precision_3
posterior_se_3 <- sqrt(1 / posterior_precision_3)

```

```

rr_prior3 <- exp(posterior_mean_3)
lower_prior3 <- exp(posterior_mean_3 - 1.96*posterior_se_3)
upper_prior3 <- exp(posterior_mean_3 + 1.96*posterior_se_3)

cat(" RR =", round(rr_prior3, 3), "\n")
cat(" 95% CrI: [", round(lower_prior3, 3), "-",
    round(upper_prior3, 3), "]\n\n")

#Prior 4: Very skeptical prior (강한 null 가정)
cat("Prior 4: Very skeptical prior (precision = 10, mean = 0)\n")
prior_precision_4 <- 10

posterior_precision_4 <- prior_precision_4 + data_precision_re
posterior_mean_4 <- (prior_precision_4 * 0 +
    data_precision_re * pooled_re) / posterior_precision_4
posterior_se_4 <- sqrt(1 / posterior_precision_4)

rr_prior4 <- exp(posterior_mean_4)
lower_prior4 <- exp(posterior_mean_4 - 1.96*posterior_se_4)
upper_prior4 <- exp(posterior_mean_4 + 1.96*posterior_se_4)

cat(" RR =", round(rr_prior4, 3), "\n")
cat(" 95% CrI: [", round(lower_prior4, 3), "-",
    round(upper_prior4, 3), "]\n\n")

cat("=== Prior Sensitivity Summary ===\n\n")

prior_results <- data.frame(
  Prior = c("Very Weak (0.1)", "Moderate (1.0)",
    "Skeptical (5.0)", "Very Skeptical (10.0)"),
  RR = c(rr_prior1, rr_bayes_base, rr_prior3, rr_prior4),

```

```

Lower = c(lower_prior1, lower_bayes_base, lower_prior3, lower_prior4),
Upper = c(upper_prior1, upper_bayes_base, upper_prior3, upper_prior4)
)

print(prior_results)

cat("\nInterpretation:\n")

cat(" - Very Weak prior: 데이터가 거의 완전히 결과 결정\n")

cat(" - Moderate prior: 균형잡힌 결과 (baseline)\n")

cat(" - Skeptical prior: null 쪽으로 shrinkage\n")

cat(" - Very Skeptical prior: null에 가장 가까움\n")

#####

##그림그리기

#####

# comparison_freq_bayes 생성
comparison_freq_bayes <- data.frame(
  Method = c("Conventional RE", "Trim-Fill", "PET", "PEESE",
             "Sel. Model S1", "Sel. Model S2", "Sel. Model S3",
             "Bayes Baseline", "Bayes S1 (2×)", "Bayes S2 (3×)", "Bayes S3
(Gradient)",
             "Prior: Very Weak", "Prior: Skeptical", "Prior: Very Skeptical"),
  RR = c(exp(m.ponv$TE.random), exp(m.trimfill$TE.random), rr_pet, rr_peese,
         exp(sel_scenario1$beta), exp(sel_scenario2$beta), exp(sel_scenario3$beta),
         rr_bayes_base, rr_bayes_s1, rr_bayes_s2, rr_bayes_s3,
         rr_prior1, rr_prior3, rr_prior4),
  Lower = c(exp(m.ponv$lower.random), exp(m.trimfill$lower.random),
            exp(log_rr_pet - 1.96*se_pet), exp(log_rr_peese - 1.96*se_peese),
            exp(sel_scenario1$ci.lb), exp(sel_scenario2$ci.lb),
            exp(sel_scenario3$ci.lb),

```

```

    lower_bayes_base, lower_bayes_s1, lower_bayes_s2, lower_bayes_s3,
    lower_prior1, lower_prior3, lower_prior4),
  Upper = c(exp(m.ponv$upper.random), exp(m.trimfill$upper.random),
    exp(log_rr_pet + 1.96*se_pet), exp(log_rr_peese + 1.96*se_peese),
    exp(sel_scenario1$ci.ub), exp(sel_scenario2$ci.ub),
exp(sel_scenario3$ci.ub),
    upper_bayes_base, upper_bayes_s1, upper_bayes_s2, upper_bayes_s3,
    upper_prior1, upper_prior3, upper_prior4),
  Type = c("Frequentist", "Frequentist", "Frequentist", "Frequentist",
    "Frequentist", "Frequentist", "Frequentist",
    "Bayesian", "Bayesian", "Bayesian", "Bayesian",
    "Bayesian", "Bayesian", "Bayesian")
)

# p_table_sorted 생성
p_table_sorted <- p_table[order(p_table$P_value), ]

# 비교 데이터 준비 (Prior sensitivity 추가)
comparison_freq_bayes$Method <- factor(comparison_freq_bayes$Method,
  levels = rev(comparison_freq_bayes$Method))

# 그림 그리기
fig_freq_bayes <- ggplot(comparison_freq_bayes, aes(x = RR, y = Method, color =
Type)) +
  geom_point(size = 3) +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.3, linewidth = 1) +
  geom_vline(xintercept = 1, linetype = "dashed", color = "darkgreen", linewidth
= 1.5) +
  scale_color_manual(values = c("Frequentist" = "steelblue", "Bayesian" = "red"))
+
  labs(title = "Frequentist vs Bayesian Approaches",
    subtitle = "Comparison of Publication Bias Adjustment Methods",
    x = "Risk Ratio (RR) with 95% CI/CrI", y = "",
    color = "Approach") +

```

```

theme_minimal(base_size = 12) +
theme(plot.title = element_text(face = "bold", size = 14),
      legend.position = "top",
      panel.grid.minor = element_blank())

ggsave("Figure6_Frequentist_vs_Bayesian.png", fig_freq_bayes,
      width = 10, height = 10, dpi = 300)
cat("Fixed order! Figure saved!\n")

#####
#

# Posterior distribution 데이터 생성
rr_seq <- seq(0.4, 1.2, length.out = 500)

# Weakly Informative (Baseline)
log_rr_seq <- log(rr_seq)
posterior_density_weak <- dnorm(log_rr_seq,
                                mean = posterior_mean_base,
                                sd = posterior_se_base)

# Very Skeptical
posterior_density_skeptical <- dnorm(log_rr_seq,
                                     mean = posterior_mean_4,
                                     sd = posterior_se_4)

# 데이터프레임 생성
posterior_df <- data.frame(
  RR = rep(rr_seq, 2),
  Density = c(posterior_density_weak, posterior_density_skeptical),
  Prior = rep(c("Weakly Informative", "Very Skeptical"), each = 500)
)

```

```
# 그림 그리기
```

```
fig_posterior <- ggplot(posterior_df, aes(x = RR, y = Density, fill = Prior)) +  
  geom_area(alpha = 0.6, position = "identity") +  
  geom_vline(xintercept = rr_bayes_base, linetype = "dashed",  
            color = "blue", linewidth = 1) +  
  geom_vline(xintercept = rr_prior4, linetype = "dashed",  
            color = "orange", linewidth = 1) +  
  geom_vline(xintercept = 1, linetype = "dashed",  
            color = "red", linewidth = 1) +  
  scale_fill_manual(values = c("Weakly Informative" = "lightblue",  
                              "Very Skeptical" = "orange")) +  
  labs(title = "C. Posterior Distributions",  
       subtitle = "Comparison of Weakly Informative vs Very Skeptical Prior",  
       x = "Risk Ratio (RR)", y = "Density") +  
  theme_minimal(base_size = 12) +  
  theme(plot.title = element_text(face = "bold", size = 13),  
        legend.position = "bottom",  
        panel.grid.minor = element_blank())  
  
ggsave("Figure_Posterior_Distributions_Panel_C.png", fig_posterior,  
       width = 8, height = 6, dpi = 300)  
cat("Panel C: Posterior Distributions saved!\n")
```

```
# =====
```

```
# P-curve 그리기
```

```
# =====
```

```
cat("=== P-value Distribution & Registry Comparison ===\n\n")
```

```
# 처음부터 다시 계산
```

```
log_rr <- log((ponv_data$event.e/ponv_data$n.e) /  
(ponv_data$event.c/ponv_data$n.c))  
se_rr <- sqrt(1/ponv_data$event.e + 1/ponv_data$event.c -
```

```

1/ponv_data$n.e - 1/ponv_data$n.c)

# P-values 계산

z_values <- log_rr / se_rr
p_values <- 2 * (1 - pnorm(abs(z_values)))

cat("P-values 계산 완료!\n")
print(round(p_values, 4))

p_table <- data.frame(
  Study = ponv_data$study_id,
  P_value = p_values,
  Significant = p_values < 0.05
)

p_table$Registry <- c(
  "None",          # Study A (p=0.008, 유의)
  "Multiplicity",  # Study B (p=0.002, 유의)
  "None",          # Study C (p=0.078, 비유의)
  "Outcome switching", # Study D (p=0.032, 유의)
  "None",          # Study E (p=0.079, 비유의)
  "Outcome switching", # Study F (p=0.006, 유의)
  "None",          # Study G (p=0.139, 비유의)
  "None",          # Study H (p=0.178, 비유의)
  "None",          # Study I (p=0.215, 비유의)
  "Multiplicity"   # Study J (p=0.021, 유의)
)

print(p_table)

```

```

# p_table_sorted 만들기
p_table_sorted <- p_table[order(p_table$P_value), ]

#확인
print(p_table_sorted)

cat("\n요약:\n")

cat("유의한 연구 (p<0.05):", sum(p_table$Significant), "개\n")
cat("  - Outcome switching:", sum(p_table$Significant & p_table$Registry ==
"Outcome switching"), "개\n")
cat("  - Multiplicity:", sum(p_table$Significant & p_table$Registry ==
"Multiplicity"), "개\n")
cat("  - None:", sum(p_table$Significant & p_table$Registry == "None"), "개\n")

cat("\n비유의 연구 (p≥0.05):", sum(!p_table$Significant), "개\n")
cat("  - None:", sum(!p_table$Significant & p_table$Registry == "None"), "개\n")

# 유의한 연구들만 (p < 0.05)
sig_studies <- p_table[p_table$Significant, ]

cat("유의한 연구:", nrow(sig_studies), "개\n")

# P-curve 히스토그램 데이터
fig_pcurve <- ggplot(sig_studies, aes(x = P_value)) +
  geom_histogram(binwidth = 0.01, fill = "steelblue", color = "black", alpha =
0.7) +
  geom_vline(xintercept = 0.025, linetype = "dashed", color = "red", linewidth =
1.5) +
  annotate("text", x = 0.035, y = max(table(cut(sig_studies$P_value, breaks =
seq(0, 0.05, 0.01)))) * 0.5,

```

```

    label = "Right-skewed:\nsuggests evidential value",
    hjust = 0, size = 4, fontface = "bold", color = "darkgreen") +
  annotate("text", x = 0.005, y = max(table(cut(sig_studies$P_value, breaks =
seq(0, 0.05, 0.01)))) * 0.8,
    label = paste0("n = ", nrow(sig_studies), " significant studies\n(p <
0.05)"),
    hjust = 0, size = 4, fontface = "bold") +
  scale_x_continuous(limits = c(0, 0.05), breaks = seq(0, 0.05, 0.01)) +
  labs(title = "A. Distribution of Significant P-values",
    subtitle = "P-curve analysis suggests evidential value (right-skewed)",
    x = "P-value", y = "Density") +
  theme_minimal(base_size = 12) +
  theme(plot.title = element_text(face = "bold", size = 13),
    panel.grid.minor = element_blank())

ggsave("Figure_7A_Pcurve.png", fig_pcurve,
  width = 8, height = 6, dpi = 300)
cat("Panel A saved!\n")

# 그림 그리기 (P-value 직접 사용)

fig_registry <- ggplot(p_table_sorted, aes(x = P_value, y = reorder(Study, -
P_value), fill = Registry)) +
  geom_bar(stat = "identity", color = "black", linewidth = 0.5) +
  geom_vline(xintercept = 0.05, linetype = "dashed", color = "red", linewidth =
1.5) +
  annotate("text", x = 0.06, y = nrow(p_table_sorted) * 0.5,
    label = "p = 0.05 threshold", hjust = 0, size = 3.5, color = "red") +
  scale_fill_manual(values = c("None" = "steelblue",
    "Outcome switching" = "orange",
    "Multiplicity" = "darkorange")) +
  scale_x_continuous(limits = c(0, max(p_table_sorted$P_value) * 1.1)) +
  labs(title = "B. Study-level Significance and Reporting Issues",
    subtitle = "Registry comparison reveals selective outcome reporting in
significant studies",
    x = "P-value", y = "Study (ordered by p-value)",

```

```
fill = "Registry Comparison") +  
theme_minimal(base_size = 12) +  
theme(plot.title = element_text(face = "bold", size = 13),  
      legend.position = "right",  
      panel.grid.minor = element_blank())  
  
ggsave("Figure_7B_Registry.png", fig_registry,  
      width = 10, height = 6, dpi = 600)  
cat("Panel B (P-value scale) saved!\n")
```

Epub ahead of print