Cureus
Part of SPRINGER NATURE

Open Access Original Article

Small-Study-Effect and Asymmetry Methods in Anatomical Prevalence Meta-Analysis: A Simulation-Informed Roadmap

Review began 10/07/2026
Review ended 10/16/2026
Published 10/20/2026

© **Copyright** 2026
Papadopoulos et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.118254

Vasileios Papadopoulos [1], George Triantafyllou [2], Maria Piagkou [2], Aliki Fiska [1]

1. Laboratory of Anatomy, School of Medicine, Democritus University of Thrace, Alexandroupolis, GRC 2. Department of Anatomy, School of Medicine, Faculty of Health Sciences, National and Kapodistrian University of Athens, Athens, GRC

Corresponding author: Vasileios Papadopoulos, vapapadop@med.duth.gr

Abstract

Anatomical prevalence meta-analyses often combine studies that differ in population, modality, specimen source, variant definition, and observational unit. Funnel-plot asymmetry is frequently interpreted as publication bias, although it may reflect transformation, event yield, heterogeneity, or prevalence-study-size associations. We evaluated six two-stage procedure families and a one-stage binomial-normal mixed-effects framework (BN-GLMM) across rare/low (0.05-5%) and broader (5-50%) prevalence domains, each with a controlled heterogeneity anchor and stochastic module. The 80,000 meta-analyses yielded 240,000 transformation-specific analyses under ordinary arcsine-square-root (PAS), Freeman-Tukey double-arcsine (PFT), and continuity-corrected two-stage logit (PLO), plus 240,000 BN-GLMM fits. Among two-stage analyses, PAS and PFT were most stable. Under PLO, heterogeneity alone generated Egger, Begg, and LFK signals even at 5-50% prevalence, whereas study-size-based regressions were more stable. BN-GLMM directly modeled counts and was highly evaluable. In rare data, all-zero evidence bases were non-estimable, and moderator estimates were unstable when only one or two studies contained events. Wald-z BN inference was well calibrated without heterogeneity but mildly anti-conservative at $K = 10$ with heterogeneity. T-reference sensitivity improved calibration and coverage with modest power loss but did not consistently restore nominal performance. Direct BN signaling was 8.8% for selection alone versus 61.8% for a generated size-prevalence association alone in rare data, and 9.4% versus 82.1% at 5-50% prevalence. Correct-direction detection declined as heterogeneity increased, estimated I^2 did not map consistently to true logit-scale heterogeneity, and selection could shift observed prevalence without detectable asymmetry. Within the evaluated assumptions, the provisional roadmap supports count-based modeling and transformation-aware triangulation. Formal inference should require $K \geq 10$, sufficient event information, and meaningful study-size variation. No asymmetry or moderator signal establishes publication bias.

Categories: Other, Anatomy
Keywords: anatomical variation, evidence-based anatomy, generalized linear mixed model, heterogeneity, meta-regression, prevalence meta-analysis, publication bias, small-study effects

Introduction

Evidence-based anatomy applies systematic evidence identification, appraisal, and synthesis to clinically relevant anatomical questions, including the prevalence and correlates of anatomical variants [1,2,3]. The underlying literature, however, is often structurally heterogeneous. Studies may be cadaveric, osteological, surgical, or imaging-based; use different detection thresholds and classification systems; and report prevalence per person, side, specimen, bone, branch, vessel, or foramen. Accordingly, a pooled estimate may be statistically precise yet have limited transportability across populations, modalities, definitions, or observational units [4,5,6].

This problem is not uncommon in clinical anatomy. A recent meta-epidemiological study of published anatomical meta-analyses reported very high heterogeneity across skeletal, neural, vascular, and other domains [7]. Comparable concerns extend beyond anatomy, because prevalence meta-analyses commonly yield extreme I^2 values that may be poorly discriminative and should not be interpreted through arbitrary thresholds [8]. Such heterogeneity may represent biological diversity, but it may also reflect ancestry, age, sex, health status, specimen preservation, imaging resolution, observer experience, anatomical landmarks, terminology, and the definition of the observational unit. Heterogeneity should therefore be interpreted as a composite feature of the evidence rather than as a single indicator of poor study quality or inappropriate pooling.

A related difficulty concerns study-level asymmetry. Doi plots and the LFK index, Egger regression, Begg-Mazumdar rank correlation, study-size meta-regression, inverse-sample-size regression, and trim-and-fill are frequently grouped under the label “publication-bias assessment." That interpretation is too strong. Plot asymmetry or an association between effect and precision may arise from genuine differences between small and large studies, for example, specialized cadaveric series versus large imaging cohorts, as well as from

between-study heterogeneity, sparse-event geometry, selective dissemination, transformation choice, methodological quality, or chance. Conversely, selective dissemination may change the pooled prevalence without generating the size-related pattern that an asymmetry procedure is designed to detect [9,10,11].

For example, suppose small osteological series use close inspection to identify tiny accessory foramina, whereas larger imaging cohorts use a higher detection threshold. The smaller studies could then report a higher prevalence even if every study were published. This hypothetical example illustrates a size-related measurement pattern, not evidence of selective dissemination.

The prevalence model or transformation can modify this geometry itself. For very low proportions, the mathematical relationship among prevalence, sampling variance, and standard error can produce funnel-plot asymmetry even when no studies are missing [10]. The Freeman-Tukey double-arcsine transformation has also generated debate regarding inversion and interpretability [12], while continuity-corrected two-stage logit analyses introduce boundary-dependent behavior; one-stage binomial mixed models avoid several of these two-stage boundary problems [13,14,15]. Thus, transformation choice is part of the asymmetry problem rather than a neutral preprocessing step.

The objectives were as follows: (1) primarily, to evaluate how commonly used small-study-effect and asymmetry procedures behave across anatomical prevalence, event-yield, transformation, and heterogeneity settings; and (2) secondarily, to describe how the true logit-scale heterogeneity used in the simulation relates empirically to estimated I^2. The goal is not to identify a universally superior method or to diagnose publication bias, but to provide a conditional, simulation-informed decision aid for anatomists and clinical researchers and to clarify what each procedure can-and cannot-support. The roadmap is provisional and has not been externally validated. This distinction is clinically relevant because pooled prevalence estimates may inform expectations in imaging, procedural planning, and variant-aware surgical practice.

Materials And Methods

Simulation architecture and study-size distributions

The investigation used two prevalence-domain platforms. The rare/low-prevalence platform generated baseline prevalences from 0.0005 to 0.05, whereas the broader-prevalence extension generated baseline prevalences from 0.05 to 0.50; sampling was log-uniform in both. Each platform contained two complementary modules: a controlled heterogeneity anchor and a stochastic simulation module. The anchor estimated method-specific background signaling when neither a genuine size-prevalence association nor selective dissemination was generated. The stochastic module combined randomly generated evidence structures to examine operating behavior under mixtures of mechanisms.

In each prevalence platform, the anchor crossed four true logit-scale heterogeneity values (τ = 0, 0.35, 0.70, and 1.20) with K = 10, 20, and 40 studies. Each of the 12 cells was replicated 2,500 times, producing 30,000 underlying meta-analyses. The stochastic module generated 10,000 underlying meta-analyses. Every underlying dataset was analyzed under PAS, PFT, and PLO and was also fitted using the three one-stage BN-GLMM specifications. Each prevalence platform therefore yielded 120,000 transformation-specific two-stage analysis sets and 120,000 one-stage fitting runs. Across both prevalence platforms, the investigation yielded 240,000 transformation-specific analysis sets and 240,000 BN-GLMM fitting runs from the same 80,000 underlying meta-analyses.

In the stochastic module, a latent mean study count (kappa) was sampled from 12, 20, 35, or 50 with equal probability (0.25 each). The actual K was generated from a truncated negative-binomial distribution with size parameter 12 and was restricted to 10-60. This design produced evidence bases with varying numbers of studies while retaining the prespecified lower limit for formal asymmetry comparisons.

For each meta-analysis, a study-size center was sampled log-uniformly from 100 to 5,000 observations and rounded to an integer. Log-scale dispersion was sampled uniformly from 0.25 to 0.90. Individual study sizes were rounded lognormal draws with meanlog = log(center) − dispersion^2/2 and sdlog = dispersion, then bounded at 30 and 10,000 by replacing out-of-range values with the corresponding limit. This created within-meta-analysis size variation and occasional large studies without imposing equal denominators.

True prevalence, heterogeneity, and generated size-prevalence associations

For study i, the true prevalence followed a logistic-normal random-effects model:

$$\operatorname{logit}(p_i) = \operatorname{logit}(p_0) + \beta_{\mathrm{SSE}} z_{\mathrm{small},i} + u_i, \qquad u_i \sim \mathcal{N}(0, \tau^2 \cdot \quad (1)$$

The direct study-size score was defined as follows:

$$z_{\text{small},i} = -\frac{\log(n_i) - \overline{\log(n)}}{\text{SD}\{\log(n)\}}. \quad (2)$$

The size score z_{small} was the negative standardized logarithm of n, so that positive values denoted smaller-than-average studies. Its mean and sample standard deviation in Equation (2) were calculated within the generated candidate pool; selected studies retained these scores unchanged for both two-stage and one-stage direct moderator analyses. The absolute genuine small-study-effect coefficient was sampled from 0, 0.25, 0.50, or 0.75 with probabilities 0.40, 0.20, 0.25, and 0.15, respectively; nonzero effects were assigned a positive or negative direction with equal probability. Study event counts were sampled from Binomial(n_i, p_i). The inverse-n moderator in Equation (7) was standardized within the retained evidence base.

The stochastic heterogeneity band was sampled as none ($\tau = 0$), low (0.10-0.30), moderate (0.30-0.60), high (0.60-1.00), or extreme (1.00-1.50), with probabilities 0.15, 0.20, 0.25, 0.25, and 0.15, respectively. Within each nonzero band, τ was sampled uniformly over its stated interval. These are true logit-scale simulation bands and should not be mapped directly to conventional I^2 categories. Estimated τ^2 and I^2 were retained from every fitted random-effects model. Because τ^2 is transformation-scale-specific and I^2 depends on within-study precision and the observed effect structure, true logit-scale τ remained the stratifying parameter.

Selective dissemination and the variance-isolated primary comparison

Candidate studies were generated before simulated dissemination. Selection type was sampled as none, zero-event, excess-event, or hybrid with probabilities 0.40, 0.15, 0.20, and 0.25. Conditional on active selection, mild, moderate, and strong settings had probabilities 0.35, 0.40, and 0.25; their score multipliers m were 0.60, 1.00, and 1.50, and target mean retention probabilities were 0.75, 0.50, and 0.30, respectively. Define $h_i = 1$ when $x_i > 0$ and 0 otherwise, and $g_i = \max(d_i, 0)$, where:

$$d_i = \frac{x_i - n_i p_0}{\sqrt{\max\{n_i p_0(1-p_0), 10^{-12}\}}}.$$

$$s_i^{\text{zero}} = m h_i, \quad s_i^{\text{excess}} = m g_i, \quad s_i^{\text{hybrid}} = m(0.90 h_i + 0.45 z_{\text{small},i} + 0.60 g_i).$$

For a candidate pool of M studies, the intercept a solved the target-retention equation below before probabilities were clipped. Each candidate was retained independently with probability r_i; under no selection, $r_i = 1$.

$$\frac{1}{M}\sum_{i=1}^{M} \text{logit}^{-1}(a + s_i) = r_{\text{target}},$$

$$r_i = \min\{0.999, \max[0.001, \text{logit}^{-1}(a + s_i)]\}.$$

The initial pool size was max{100, ceiling(Kc)}, where c = 1.2 without selection and c = max{2, 2.5/target} otherwise. If fewer than K studies were retained, a fresh pool of twice the size was generated and recalibrated, for up to 10 attempts. Once sufficient studies were retained, K were sampled uniformly without replacement. Clipping and realized Bernoulli retention can make the observed retention fraction differ from the target. The rare-domain code used a root bracket of −30 to 30; the broader-domain code used the score-adaptive bracket documented in the archived source.

This mechanism was deliberately stylized. It was designed to distinguish selective dissemination from a genuine study-size effect, not to reproduce a literal editorial, investigator, or journal selection process. Four data-generating mechanism groups resulted: neither mechanism, genuine small-study effect only, selection only, and both mechanisms.

The primary comparative simulations fixed the generic variance-inflation factor at 1.00. In the separate earlier sensitivity, a factor F was sampled as 1.00, 1.25, 1.50, or 2.00 with probabilities 0.55, 0.20, 0.15, and 0.10. After counts were transformed, independent standard-normal perturbations Z_i were added as follows, using the same Z_i across transformations t while leaving each reported variance $v_{i,t}$ unchanged:

$$y^{\text{pert}}{}_{i,t} = y_{i,t} + \sqrt{(F-1)v_{i,t}}\, Z_i, \qquad Z_i \sim N(0,1).$$

The superscript "pert" denotes the perturbed transformed effect. The stochastic module restricted to neither a generated size association nor selection is summarized in the Results section. This is a generic

transformed-effect variance stress test, not an explicit bilateral, clustered-binomial, or beta-binomial data-generating model. It did not determine roadmap ratings, and its outputs are separate from the 80,000 primary evidence bases.

Two-stage transformations and random-effects synthesis

Each dataset was analyzed under PAS, PFT, and PLO. PAS used the ordinary arcsine-square-root transformation without continuity correction. PLO used a logit-transformed proportion with a 0.5 correction only in boundary studies (zero events or events in the full sample). PFT used the Freeman-Tukey double-arcsine transformation and a harmonic-mean-based inverse transformation. Random-effects models were fitted using restricted maximum likelihood. Applying all three transformations to each dataset created a matched design in which disagreement could be attributed to the analytical scale rather than to different simulated evidence bases.

PFT was included because it remains common in anatomical prevalence meta-analysis, not because the simulation establishes it as the preferred estimator. The literature contains competing positions regarding its practical performance and inverse transformation [12,13,15]. These transformation-based analyses were complemented by a separate one-stage binomial-normal mixed-effects framework that modeled the original event counts conditional on their study denominators. The BN-GLMM was treated as a distinct likelihood-based model family rather than as a fourth transformation. Beta-binomial, Bayesian, multilevel, cluster-sensitive, and alternative random-effects-distribution models were not evaluated.

One-stage binomial-normal mixed-effects models

For study i, let X_i denote the observed number of anatomical events among n_i observations. Conditional on a study-specific random effect u_i, the one-stage binomial-normal generalized linear mixed model used a binomial likelihood with normally distributed study effects on the logit scale, following the likelihood framework described by Stijnen et al. [16].

$$X_i \mid u_i \sim \mathrm{Binomial}(n_i, p_i), \qquad u_i \sim \mathcal{N}(0, \tau^2). \quad (3)$$

Three model specifications were fitted to each underlying evidence base.

The intercept-only model provided a count-based synthesis benchmark and an estimate of logit-scale heterogeneity:

$$\mathrm{logit}(p_i) = \beta_0 + u_i. \quad (4)$$

The direct study-size moderator model evaluated the standardized size construct used in the data-generating model:

$$\mathrm{logit}(p_i) = \beta_0 + \beta_{\mathrm{size}} z_{\mathrm{small},i} + u_i. \quad (5)$$

Positive values of $z_{small,i}$ denoted smaller-than-average studies. The generated coefficient β_{SSE} and the fitted direct moderator coefficient β_{size} therefore referred to the same size-related construct.

The inverse-sample-size model evaluated an indirect proxy for the study size:

$$\mathrm{logit}(p_i) = \beta_0 + \beta_{\mathrm{inv}} z_{\mathrm{inv},i} + u_i. \quad (6)$$

$$z_{\mathrm{inv},i} = \frac{n_i^{-1} - \overline{n^{-1}}}{\mathrm{SD}(n^{-1})}. \quad (7)$$

Models were fitted by maximum likelihood using metafor::rma.glmm(), the lme4 integration engine, internal predictor scaling, and adaptive Gaussian quadrature with seven nodes. Primary inference used two-sided Wald-z tests. In the rma.glmm() function, the argument measure = "PLO" selected the logit-proportion parameterization of this binomial model. The one-stage likelihood was fitted directly to the observed event counts and study denominators, not to precomputed logit-transformed proportions and estimated sampling variances. The 0.5 boundary correction affected only auxiliary study-level effect and variance quantities retained in the fitted object; it did not modify the counts used in the likelihood. Zero-event studies were therefore retained without continuity correction in one-stage likelihood estimation.

A model was not fitted when fewer than 10 studies were available, when the complete evidence base contained no events, when all observations were events, or when the specified moderator had effectively no variation. A returned fit was classified as estimable when the intercept and a nonnegative estimate of τ^2

were finite and, for a moderator model, the moderator coefficient, its positive standard error, and its p-value were also finite. A clean fit was an estimable fit without a standard-error flag or captured fitting warning. Model errors, nonfinite returns, standard-error warnings, and other fitting warnings were recorded separately.

Information-state summaries included total events and non-events; the numbers of event-bearing, non-event-bearing, zero-event, and full-event studies; and classification of each evidence base as all zero, all full, containing one event-bearing study, containing two event-bearing studies, or containing at least three event-bearing studies. These categories described the information gradient and were not used to impose a universal eligibility threshold.

For the moderator models, the primary operational signal was a two-sided Wald-z p-value below 0.05. Correct-direction detection required both $p < 0.05$ and agreement between the signs of the estimated and generated coefficients; it was defined only when β_{SSE} was nonzero. Coefficient bias, squared error, and 95% confidence-interval coverage were recorded for the direct study-size coefficient. Intercept and heterogeneity recovery were evaluated on the logit scale. The inverse-logit intercept represented conditional prevalence at moderator value zero and random effect $u_i = 0$; it was not interpreted as the marginal arithmetic mean prevalence across studies.

Small-study-effect and asymmetry procedures

Within each two-stage transformation, six procedure families were evaluated. Direct mixed-effects study-size meta-regression modeled the transformed effect as a function of zsmall and addressed whether prevalence was systematically related to study size. Inverse-sample-size regression used 1/n as an indirect, size-based comparator that was less directly tied to the estimated standard error. Egger et al.'s regression [17] used the standard error as its predictor and evaluated an effect-precision association without identifying its cause. The Begg-Mazumdar procedure [18] assessed rank association between the effect and variance/precision and served as a supplementary rank-based assessment. The Doi plot and LFK index of Furuya-Kanamori et al. [19] described the magnitude and direction of asymmetry; LFK was recorded continuously and at the legacy $|LFK| > 2$ threshold. Duval and Tweedie's trim-and-fill method [20] used rank-based imputation of putatively missing studies and was treated as a counterfactual sensitivity adjustment, not a diagnostic test.

A method-specific signal was defined as $p < 0.05$ for moderator, regression, and rank procedures; $|LFK| > 2$ for the legacy LFK criterion; and $k0 > 0$ for descriptive trim-and-fill output. The intercept-only BN-GLMM did not generate an asymmetry signal; its recorded outcomes were estimability, fitting warnings, clean-fit status, the conditional intercept, and heterogeneity recovery. A signal indicates only that an operational criterion was met: it does not establish publication bias or identify the cause of asymmetry.

For the one-stage moderator models, formal inference required at least 10 studies, at least one event and one non-event across the complete evidence base, and meaningful moderator variation. For the two-stage procedures, formal p-value-based analysis required at least 10 analyzable studies and variation in both effects and the relevant study-size or precision-related quantity. These conditions defined the operating domain evaluated here and did not guarantee adequate power. The LFK index was computed whenever at least two finite transformed effects and standard errors were available, but conventional LFK categories were not validated below $K = 10$. Trim-and-fill was attempted with at least three analyzable studies but was excluded from primary diagnostic-accuracy summaries. A model-based or asymmetry signal indicates only that the procedure's operational criterion was met; it does not identify the cause of the pattern or establish publication bias.

The prespecified function lfk_index_manual() sorted finite transformed effects y_i and positive standard errors s_i by y_i. With $v_i = s_i^2$ and k retained studies, it calculated precision-weighted pseudo-counts N_i, weighted midranks H_i, and normal scores z_i:

$$N_i = 100\frac{\max_j v_j}{v_i}, \qquad H_i = \sum_{\ell<i} N_\ell + \frac{N_i}{2},$$

$$z_i = \Phi^{-1}\left(\frac{H_i - 0.5}{\sum_{\ell=1}^{k} N_\ell}\right),$$

Writing y_j for the effect whose $|z_i|$ was smallest, the index was:

$$\mathrm{LFK} = \frac{5}{2k}\sum_{i=1}^{k}\left(z_i + \frac{\max_\ell z_\ell - \min_\ell z_\ell}{\max_\ell y_\ell - \min_\ell y_\ell}(y_i - y_j)\right).$$

The implementation clipped percentile arguments to [10^{-12}, 1 – 10^{-12}], returned an unavailable value with fewer than two valid pairs, and returned zero for an effect range below 10^{-12}. These are explicit numerical conventions. During revision, independent Python evaluations of the manual and metasens-source arithmetic agreed within 10^{-10} for 1,001 nondegenerate fixtures, including the published pain example. Guard behavior was checked separately. This arithmetic audit is not execution of the R package or validation of the legacy threshold; the executable R-to-metasens comparison script and audit details are included in the reproducibility materials.

Event-yield strata and performance outcomes

The rare/low-prevalence analysis summarized sparse information using median expected events per study, categorized as ⩽0.25, 0.25-0.75, 0.75-1.5, 1.5-5, 5-15, 15-50, or >50. The broader-prevalence extension used widened categories: ⩽5, 5-15, 15-50, 50-150, 150-500, 500-1500, or >1500 expected events per study. It also grouped baseline prevalence as 5-10%, 10-20%, 20-30%, 30-40%, or 40-50%. The observed fraction of zero-event studies was retained in both platforms.

The primary outcome for moderator, regression, rank, and geometric procedures was method-specific signal probability among evaluable simulations. In the controlled anchor, this represented background signaling because neither a genuine size-prevalence association nor selective dissemination was generated. In the stochastic module, signal probability was summarized separately for evidence bases containing neither mechanism, selection only, a genuine size-prevalence association only, or both mechanisms.

For the one-stage models, additional outcomes included structural non-estimability, model errors, standard-error or other fitting warnings, clean-fit probability, total event yield, and the number of event-bearing studies. Direct study-size moderator performance included coefficient bias, squared error, 95% interval coverage, and correct-direction detection. The corresponding Wald-z outcomes were compared with the derived t-reference sensitivity using K-2 degrees of freedom. Intercept-only BN-GLMM estimates were used to assess recovery of the generated logit-scale τ.

The direct BN-GLMM moderator was paired with each PAS, PFT, and PLO direct study-size meta-regression fitted to the same underlying evidence base. Paired outcomes included overall binary classification agreement, BN-only signaling, two-stage-only signaling, and discordance. Numerical coefficient differences were evaluated only between BN-GLMM and PLO because both coefficients were expressed on the logit scale; PAS and PFT coefficients were not compared numerically across incompatible transformation scales. Logit-scale heterogeneity recovery was compared between the intercept-only BN-GLMM and two-stage PLO in the controlled anchor and in stochastic evidence bases without a generated size effect or selection.

Method-specific logistic signal-driver models used signal status as the dependent variable and included the natural logarithm of the number of studies, the natural logarithm of one plus the median expected number of events per study, standardized true logit-scale heterogeneity, standardized within-meta-analysis variability in log-transformed study size, the presence of a generated size-prevalence association, and the presence of selective dissemination as predictors. Continuous covariates were standardized using the unique underlying stochastic evidence bases within each prevalence domain before expansion by analytical method.

Additional outcomes retained from the original simulation programme included transformation disagreement on the same evidence base, retention rate, selection-induced shifts in true and observed prevalence, pooled-prevalence bias, and the finite-sample rare-event reporting-bias calculations. A method-specific signal denotes that the operational criterion was met and should not be interpreted as a diagnosis of publication bias.

Derived t-reference sensitivity

Small-K inference was examined by recalculating moderator p-values and confidence intervals from stored coefficient estimates and standard errors. For q fixed-effect coefficients, the sensitivity used the following statistic, two-sided reference probability, and 95% confidence interval:

$$T = \frac{\widehat{\beta}}{\mathrm{SE}(\widehat{\beta})}, \qquad \nu = K - q.) \quad (8)$$

$$p_t = 2\left[1 - F_{t_\nu}\left(|T|\right)\right]. \quad (9)$$

$$\mathrm{CI}_{0.95} = \widehat{\beta} \ \pm \ t_{\nu,0.975}\mathrm{SE}(\widehat{\beta}). \quad (10)$$

Both moderator models contained an intercept and one moderator, so q = 2 and the reference degrees of freedom were K − 2. This was a post-processing sensitivity; neither the coefficients nor their standard errors were re-estimated.

Empirical crosswalk between true heterogeneity and estimated I^2

Because the intended applied audience commonly encounters I^2, a secondary translational analysis used the replicate-level outputs already produced by both simulation platforms. Within the controlled anchors, finite I^2 values were summarized separately by prevalence domain, transformation, K, and true τ using the median, interquartile range, 5th-95th percentiles, and proportions in descriptive bands (<25%, 25%-<50%, 50%-<75%, and ⩾75%). These bands were not used to derive or alter the decision-aid ratings. The stochastic modules were also restricted to true $\tau = 0$ and summarized by the generated mechanism to evaluate whether a genuine size-prevalence association or selective reporting could contribute to observed inconsistency without random between-study dispersion. Direct comparison of estimated τ across PAS, PFT, and PLO was not attempted because random-effects heterogeneity is expressed on the scale of the fitted transformation. A separate comparison was nevertheless conducted between the intercept-only BN-GLMM and two-stage PLO because both estimated heterogeneity on the logit scale. This comparison supplemented, but did not replace, the empirical I^2 crosswalk.

Finite-sample rare-event reporting-bias functional

The rare/low-prevalence platform additionally evaluated uncertainty for a finite-sample selective-reporting-bias functional previously proposed for rare anatomical variants. For an observed count x among n observations, define $\hat{p} = x/n$ and the following quantities [21]:

$$q_n(x) = \frac{V}{V+W}, \qquad V = \Pr_{\widehat{p}}(X=0), \qquad W = \Pr_{\widehat{p}}(X=x). \quad (11)$$

$$b_n(x) = \widehat{p}\, q_n(x), \qquad \widehat{p} = \frac{x}{n}. \quad (12)$$

Exact finite-sample moments of $b_n(x)$ were calculated under the unconditional binomial distribution, conditional on $X > 0$, and under the simulated selection-weighted event-count distribution. These exact SDs were compared with a fixed-fraction delta approximation and a naïve binomial-proportion approximation. The 5-50% extension disabled this calculation by default because the functional targets low-event omission, becomes computationally intensive at higher event counts, and was secondary to the broad-prevalence question.

Construction and interpretation of the roadmap

The roadmap integrated four simulation-based criteria: background calibration in the controlled heterogeneity anchor; responsiveness to a generated size-prevalence association; robustness across transformation and event yield; and each method's legitimate inferential role. It is a pragmatic decision aid, not a validated clinical classification model. For the regression family, correct-direction detection below 50% for the smallest generated effect, $|\beta SSE| = 0.25$, was used as a practical reason to downgrade a prevalence/heterogeneity stratum from primary-candidate to adjunct use, even when background signaling was comparatively favorable. The 50% boundary was a pragmatic flag for detecting the smallest imposed effect in fewer than half of the evaluated simulations, not a conventional adequate-power target, an optimized cut-point, or an externally validated decision threshold. In this revision, the detection criterion was examined with 95% Monte Carlo intervals and alternative 40% and 60% floors. These are sensitivity descriptions, not new automatic rules for assigning all P/A/S ratings. A rating near a boundary is provisional and should be interpreted as adjunct-level evidence when its Monte Carlo interval overlaps that boundary.

P (primary candidate) indicated comparatively favorable background signaling, a direct or closely related size-based question, and sufficient information in the tested stratum; it did not certify a 5% false-positive rate or 95% coverage in every cell. Such a procedure could serve as a conditional primary candidate for its stated role, subject to model assumptions, the reported calibration limitations, Monte Carlo uncertainty, and causal caution. An (adjunct) indicated reduced power, noninvariant thresholds, analytical-scale sensitivity, or an informative but indirect construct; it called for triangulation, explicit limitations, and interpretation of discordance. S (sensitivity/descriptive only) indicated frequent signals under null/heterogeneity conditions or a counterfactual adjustment algorithm rather than a diagnostic procedure. An S-rated result should not be the principal basis for concluding publication bias or excluding a genuine size-related pattern.

The heterogeneity contexts correspond to the known true logit-scale τ in the simulation: none/low ($\tau = 0$ or 0.10-0.30), moderate (0.30-0.60), and high/extreme (0.60-1.50). These are simulation strata, not categories that can be recovered mechanically from empirical I^2.

The roadmap assumes $K \geqslant 10$ and meaningful variation in study size or precision. With $K < 10$, the P/A/S ratings do not apply: classify the evidence as insufficient for formal asymmetry assessment, while considering pooling separately on anatomical and methodological grounds. K = 10 is an applicability boundary, not a guarantee of adequate power.

The roadmap applies only after the observational unit and dependence structure have been defined correctly. No asymmetry tool repairs bilateral or clustered-data misspecification.

These categories concern the tested analysis scale and target question. A P rating does not denote a publication-bias test; an A rating does not mean that calculation is invalid; and an S rating does not mean that calculation is prohibited. PLO-based results may be reported with a more robust scale and an explicit limitation. PFT and PLO ratings do not endorse either transformation as the preferred pooled-prevalence estimator.

Software, Monte Carlo precision, validation, and reproducibility

The simulations were conducted in R 4.6.1 using the metafor package described by Viechtbauer [22] and supporting packages. Principal packages and versions were metafor 5.0-1, lme4 2.0-6, numDeriv 2016.8-1.1, dplyr 1.2.1, tidyr 1.3.2, purrr 1.2.2, readr 2.2.0, and tibble 3.3.1. The prespecified simulation seed was 20260730. Seeds were generated using an overflow-safe modular mapping, and the controlled-anchor and stochastic runs were stored in restartable checkpoint chunks.

The rare/low-prevalence simulations fixed the generic variance-inflation factor at exactly 1.00 and verified this condition before summaries were generated. For the 5-50% prevalence simulations, a score-adaptive root bracket was used to prevent selection-intercept calibration failures while preserving the target-retention objective. The complete workflow regenerated all 80,000 underlying evidence bases, verified their exact correspondence with the embedded two-stage reference outputs, fitted the three BN-GLMM specifications, and integrated the one-stage and two-stage results by prevalence domain, simulation module, and replicate identifier. The derived t-reference sensitivity used the stored BN coefficient estimates and standard errors and therefore did not require model refitting.

For an estimated signal probability $\hat{p}$ based on R evaluable simulation replicates, Monte Carlo precision was summarized as follows:

$$\mathrm{MCSE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{R}}. \quad (13)$$

For the revision, Monte Carlo uncertainty was recalculated directly from stored replicate-level outcomes using each method-specific evaluable denominator. Two-sided 95% Wilson intervals describe uncertainty in simulated probabilities, not uncertainty in an anatomical prevalence estimate. The accompanying audit reports successes, denominators, MCSEs, and intervals for $|\beta_{SSE}| = 0.25$ by individual heterogeneity band and pooled context, with no-selection and all-selection-state subsets kept separate. It also reports BN-GLMM anchor and mechanism-specific signaling intervals. No evidence bases were regenerated, and no models were refitted for these additional summaries.

The reproducibility materials comprising the portable package Clinical_Anatomy_Portable_Reproducibility and the revision validation_addendum can be found in the Appendix. The portable package contains readable R sources under source/ and R/, dependency manifests and installation instructions under environment/, session information and frozen replicate-level outputs under reference/, and archived exact-reproduction reports. Its run_all_analyses.R entry point provides validation, smoke-test, and full-reconstruction routes. The revision addendum contains the Monte Carlo audit, selection/LFK/variance-inflation implementation notes, arithmetic-validation results, and the worked-example extract. Access details must be supplied in the Data Availability statement. The archived full-run validation and the revision's independent downstream checks are distinguished from a new execution of the complete R workflow.

Use of generative artificial intelligence for language editing and R-code debugging is disclosed in the Acknowledgments. The authors retained responsibility for the study design, analytical decisions, interpretation, and final text.

Results

Findings common to both prevalence domains

Across both prevalence domains, the principal two-stage hierarchy was consistent. Under PAS and PFT, direct study-size meta-regression, inverse-sample-size regression, and Egger-SE generally showed low-to-moderate background signaling and responded when a genuine relationship between study size and underlying prevalence was generated. Two-stage direct study-size meta-regression was most closely aligned with the simulated scientific question because study size itself was the moderator. Inverse-sample-size regression was similarly stable across transformations, whereas Egger-SE was more dependent on the transformation-specific relationship between effect and standard error.

Doi/LFK remained useful for describing the magnitude and direction of asymmetry, but exceedance of the

legacy |LFK| > 2 threshold varied with transformation, event yield, heterogeneity, and study count [23]. Begg's test was less discriminating and was particularly reactive in extremely sparse rare-event strata. Trim-and-fill frequently imputed one or more studies even when no selection mechanism was generated, supporting its use as a counterfactual sensitivity analysis rather than as evidence that unpublished studies exist.

Rare and low baseline prevalence: simulation yield and evaluability

The rare/low-prevalence analysis generated 40,000 underlying meta-analyses: 30,000 in the controlled heterogeneity anchor and 10,000 in the stochastic simulation module. Matched PAS, PFT, and PLO analysis yielded 120,000 transformation-specific analyses. In the stochastic module, the median K was 25 (IQR 16-39; range 10-60). The median of within-meta-analysis median study sizes was 602 observations (IQR 227-1,587), the median baseline prevalence was 0.50% (IQR 0.16-1.59%), and the median expected event yield was 3.24 events per study (IQR 0.93-12.40). The median zero-event fraction was 9.1% (IQR 0-34.5%). Of the 10,000 stochastic meta-analyses, 1,633 contained neither mechanism, 2,280 contained a genuine small-study effect only, 2,450 contained selection only, and 3,637 contained both mechanisms.

LFK was computable in all transformation-specific analyses. PAS/PFT evaluability of formal regression and rank procedures ranged from 98.5% to 99.7%. PLO evaluability was lower, ranging from 93.2% for Egger to 96.0% for Begg; trim-and-fill was evaluable in 92.2% of PLO analyses.

Rare and low prevalence: controlled heterogeneity anchor

Under PAS, Egger signal probabilities ranged from 0.068 to 0.123 across the 12 K-by-τ conditions. The corresponding ranges were 0.065-0.121 for inverse-n regression and 0.066-0.116 for direct study-size meta-regression. Legacy |LFK| > 2 occurred in 0.070-0.184, while Begg signals ranged from 0.093 to 0.210. PFT produced a similar profile: Egger 0.036-0.092, direct study-size meta-regression 0.036-0.094, inverse-n regression 0.034-0.077, and LFK threshold exceedance 0.069-0.175; Begg reached 0.242-0.255 in the K = 40 strata (Figure *1*).

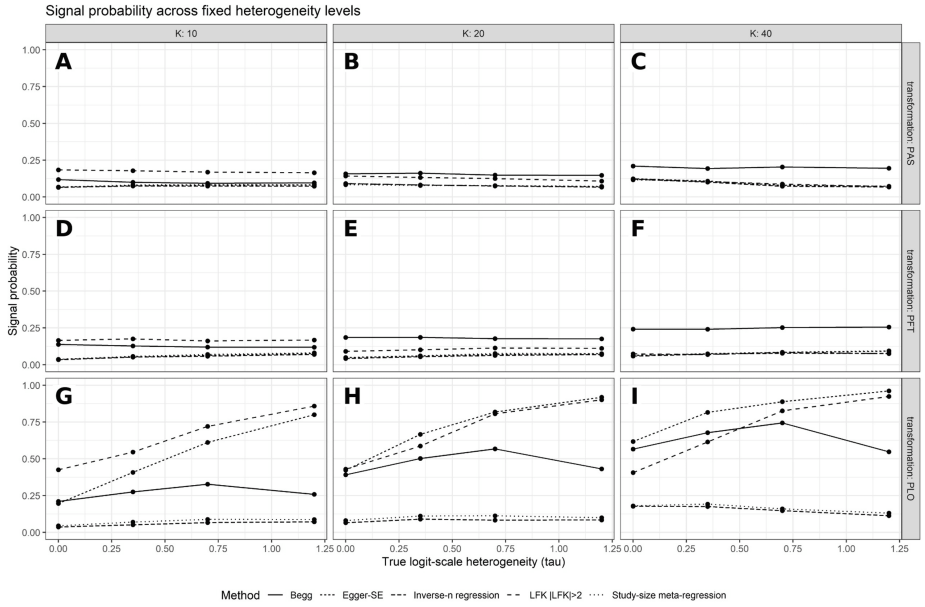


FIGURE 1: Rare/low-prevalence controlled heterogeneity anchor.

Signal probabilities across true τ = 0, 0.35, 0.70, and 1.20. Panels A/B/C show PAS with K = 10/20/40; D/E/F show PFT with K = 10/20/40; G/H/I show PLO with K = 10/20/40. No genuine size association or selection was generated. A signal was $p < 0.05$ for formal procedures or |LFK| > 2 for the legacy LFK threshold.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

PLO had a different operating surface. Egger signaling increased from 0.196 at K = 10 and τ = 0 to 0.962 at K = 40 and τ = 1.20. LFK threshold exceedance increased from 0.426 to 0.924 across the same conditions, and Begg reached 0.744. Inverse-n and direct study-size meta-regression remained much less reactive, with maximum signal probabilities of 0.177 and 0.193. These are background signals because no genuine size effect or selection was generated.

Figure 1A, 1B, and 1C show PAS at K = 10, 20, and 40, respectively; Figure 1D, 1E, and 1F show the corresponding PFT analyses; and Figure 1G, 1H, and 1I show PLO. The lower row makes the heterogeneity-related increase in PLO-based Egger and LFK signaling visible, in contrast to the more stable study-size-based regressions.

Rare and low prevalence: transformation disagreement

Among datasets with all three transformation-specific results evaluable, at least one discordant classification occurred in 72.8% for LFK, 55.4% for Egger, 50.3% for trim-and-fill, and 45.8% for Begg. Direct study-size meta-regression and inverse-n regression were more stable, with disagreement rates of 18.9% and 18.0%. LFK pairwise disagreement was 12.9% between PAS and PFT, 64.2% between PAS and PLO, and 68.5% between PFT and PLO.

Rare and low prevalence: generated size effects versus selection

Under PAS, Egger signaled in 8.8% of datasets containing neither mechanism, 8.5% with selection only, 58.9% with a genuine small-study effect only, and 61.2% with both. Direct study-size meta-regression yielded 9.4%, 8.6%, 59.0%, and 61.4%; inverse-n regression yielded 8.2%, 8.5%, 57.2%, and 59.5% (Figure 2). PAS LFK threshold exceedance was 13.0%, 10.3%, 33.9%, and 33.0%, while Begg yielded 17.1%, 12.9%, 56.5%, and 54.4%.

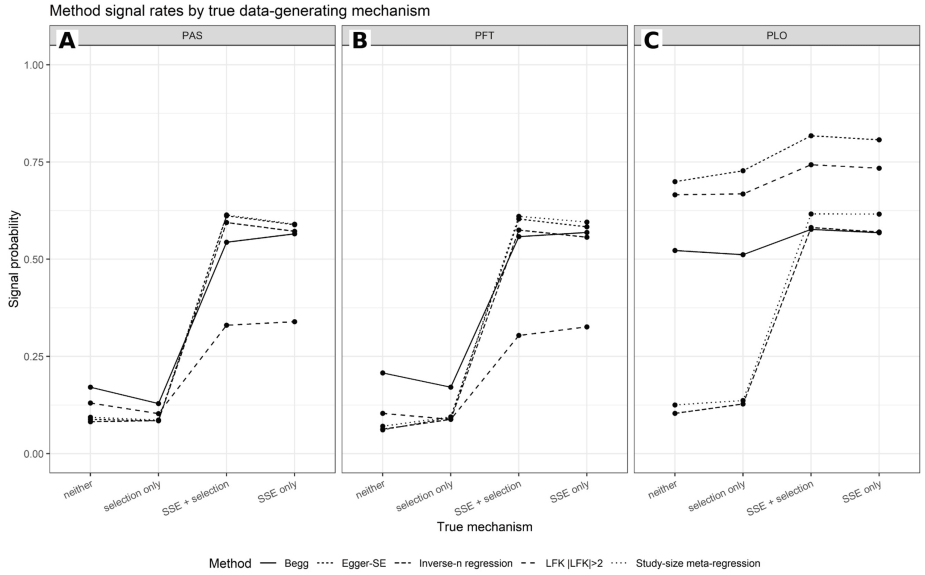


FIGURE 2: Rare/low-prevalence signals by data-generating mechanism.

Panel A: PAS; B: PFT; C: PLO. From left to right, the four states are neither mechanism, selection only, both mechanisms, and SSE only. Rates are proportions among evaluable stochastic analyses. SSE denotes a relationship between true prevalence and standardized study size.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

PFT was closely comparable. Egger signal probabilities were 6.1%, 9.3%, 58.3%, and 60.3% across neither, selection only, SSE only, and both. Direct study-size meta-regression yielded 7.0%, 9.5%, 59.6%, and 61.0%, while LFK threshold exceedance yielded 10.3%, 8.8%, 32.6%, and 30.4%.

PLO showed high background signaling: in the neither condition, Egger, LFK, and Begg signaled in 69.9%, 66.6%, and 52.2%, respectively. Under an SSE only, the corresponding probabilities were 80.7%, 73.4%, and 56.8%. Direct study-size meta-regression increased from 12.5% in neither to 61.6% with an SSE, and inverse-n regression increased from 10.4% to 57.0%.

Selection was active despite limited additional PAS/PFT asymmetry. Under strong selection, mean retention rates were 29.9% for excess-event, 30.1% for hybrid, and 30.3% for zero-event selection. Mean observed-prevalence shifts between selected and candidate pools were +1.10, +1.36, and +0.09 percentage points, respectively. Thus, selective dissemination could materially change the observed prevalence without necessarily creating a detectable size-prevalence association.

Trim-and-fill imputed at least one study in 53.8% of PAS and 49.6% of PFT datasets containing neither mechanism; under selection only, the corresponding rates were 51.4% and 54.0%. Under PLO, one or more studies were imputed in 92.0% of neither and 88.3% of selection-only datasets, with a median k0 of seven in both conditions.

Figure 2A (PAS), Figure 2B (PFT), and Figure 2C (PLO) contrast the four generated mechanism states. The PAS and PFT panels distinguish a generated size-related pattern from selection alone, whereas the PLO panel also displays substantial background signaling for effect-precision, rank, and geometric procedures.

Rare and low prevalence: event yield and heterogeneity

Figure 3 summarizes the joint event-yield-by-heterogeneity surface across all four data-generating mechanisms. Under PAS/PFT, signal probabilities of Egger, inverse-n regression, and direct study-size meta-regression generally increased with event yield when heterogeneity was absent or low. Under PAS with no heterogeneity, study-size meta-regression increased from 0.195 at ⩽0.25 expected events to 0.660 at >50 events; under low heterogeneity, it increased from 0.195 to 0.694. Under extreme heterogeneity, its signal probability was only 0.091-0.283 across event-yield strata.

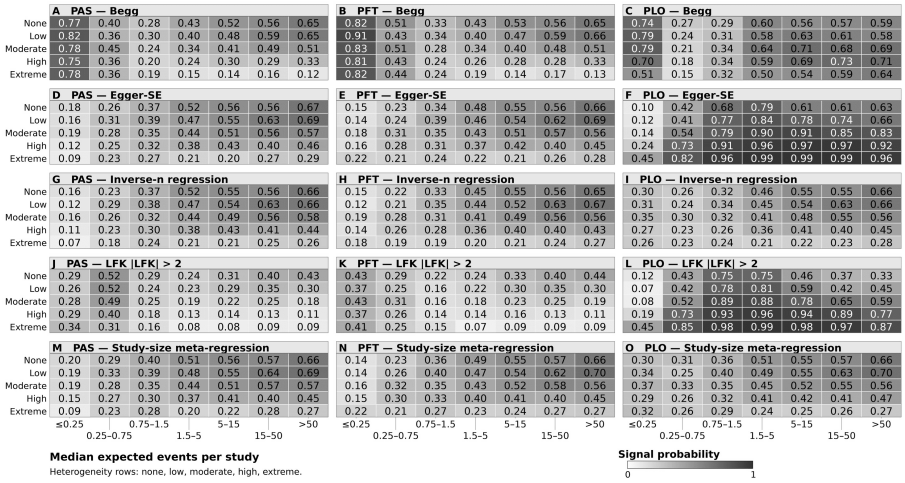


FIGURE 3: Rare/low-prevalence signal probability by event yield and heterogeneity.

Panels A/B/C: Begg; D/E/F: Egger-SE; G/H/I: inverse-n regression; J/K/L: LFK; M/N/O: direct study-size meta-regression. Within every triplet, panels are PAS, PFT, and PLO, respectively. Rows run from none to extreme true logit-scale heterogeneity; columns within panels are median expected-event strata. Values combine all four mechanism states and are not pure null or power estimates.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

Begg was unusually reactive at the extreme sparse boundary: with ⩽0.25 expected events, signal probabilities ranged from 0.747 to 0.819 across PAS heterogeneity bands and from 0.810 to 0.907 across PFT bands. LFK was nonmonotonic. Under PAS with no/low heterogeneity, the highest threshold-exceedance probabilities occurred at 0.25-0.75 expected events (0.517 and 0.523). PLO-based Egger and LFK remained high over broad portions of the surface; in extreme heterogeneity, both reached approximately 0.96-0.99 in several event-yield strata.

In Figure 3, panels A/B/C show Begg, D/E/F Egger-SE, G/H/I inverse-n regression, J/K/L LFK, and M/N/O direct study-size meta-regression; within each triplet, the order is PAS, PFT, and PLO. Rows within each panel run from none to extreme true heterogeneity. These panels summarize all four generated mechanism states and therefore describe an overall operating surface rather than pure null calibration or power.

Rare and low prevalence: correct-direction detection

Correct-direction detection increased with $|\beta_{SSE}|$ and decreased with heterogeneity (Figure 4).

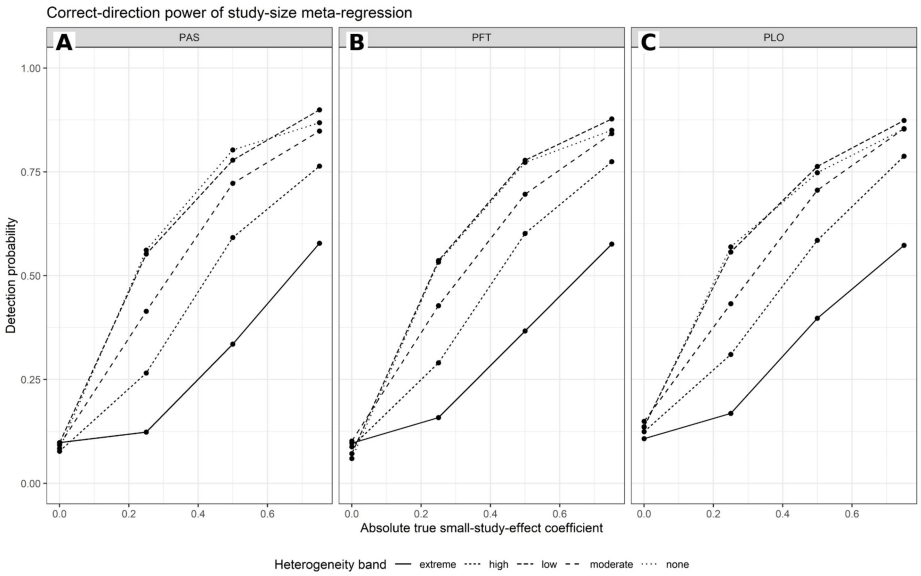


FIGURE 4: Rare/low-prevalence correct-direction detection by study-size meta-regression.

Panel A: PAS; B: PFT; C: PLO. Detection required $p < 0.05$ and a coefficient with the same sign as the generated SSE. At $|\beta SSE| = 0$, values represent two-sided background signal rates rather than directional detection.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

At $\beta_{SSE} = 0$, empirical two-sided signal probabilities were approximately 0.06-0.15. For $|\beta_{SSE}| = 0.25$, PAS detection was 0.562 with no heterogeneity, 0.552 under low, 0.414 under moderate, 0.266 under high, and 0.123 under extreme heterogeneity. PFT yielded 0.533, 0.536, 0.428, 0.290, and 0.158; PLO yielded 0.569, 0.557, 0.433, 0.310, and 0.168. At $|\beta_{SSE}| = 0.50$, detection was approximately 0.75-0.80 under none/low but 0.335-0.398 under extreme heterogeneity. At $|\beta_{SSE}| = 0.75$, detection was approximately 0.85-0.90 under none/low and 0.57-0.58 under extreme heterogeneity.

Figure *4A* (PAS), Figure *4B* (PFT), and Figure *4C* (PLO) show how correct-direction detection increases with the absolute generated coefficient and decreases with heterogeneity. Values at a coefficient of zero are two-sided background signal probabilities, not directional power. A non-significant moderator result cannot exclude a genuine size-related pattern when heterogeneity is substantial.

Rare-event reporting-bias uncertainty and legacy variance-inflation sensitivity

The fixed-fraction approximation had a median approximate-to-exact SD ratio of 0.717-0.913 across selection mechanisms. Only 34.9-36.8% of values were within a factor of two of the selection-weighted exact SD; 34.9-39.7% were less than one-half, and 25.3-30.1% exceeded twice the exact SD. The naïve binomial approximation was more dispersed: median ratios 0.523-0.813; only 23.3-26.0% within a factor of two; 44.7-49.4% less than one-half; and 26.5-32.1% above twice the exact SD (Figure *5*).

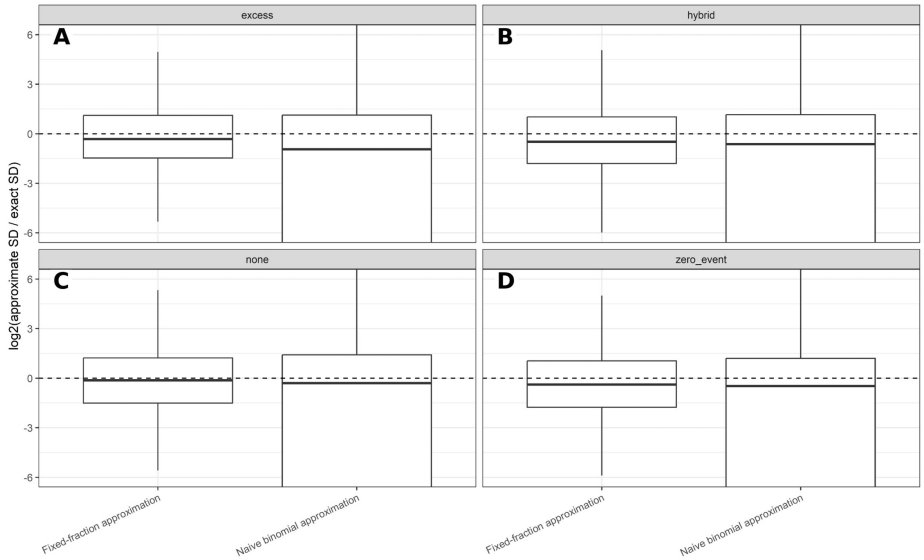


FIGURE 5: Approximate versus exact selection-weighted reporting-bias standard deviation.

Panel A: excess-event selection; B: hybrid selection; C: no selection; D: zero-event selection. Base-2 logarithm of the approximate SD divided by the exact finite-sample selection-weighted SD for the rare-event reporting-bias functional. Zero denotes equality; +1 denotes twofold overestimation and −1 denotes twofold underestimation. Distributions compare the fixed-fraction delta and naïve binomial approximations across selection mechanisms.

Figure *5A* compares approximations under excess-event selection, Figure *5B* under hybrid selection, Figure *5C* with no selection, and Figure *5D* under zero-event selection. Each panel contrasts the fixed-fraction delta and naïve binomial approximations against the exact finite-sample standard deviation.

Figure *6* retains the earlier variance-inflation sensitivity as a separate legacy analysis. Panels A (PAS), B (PFT), and C (PLO) illustrate how unaccounted transformed-effect variance can alter apparent signal rates. The primary variance-isolated simulations fixed the variance-inflation factor at 1.00; this earlier version 2.3 sensitivity did not determine any roadmap rating.

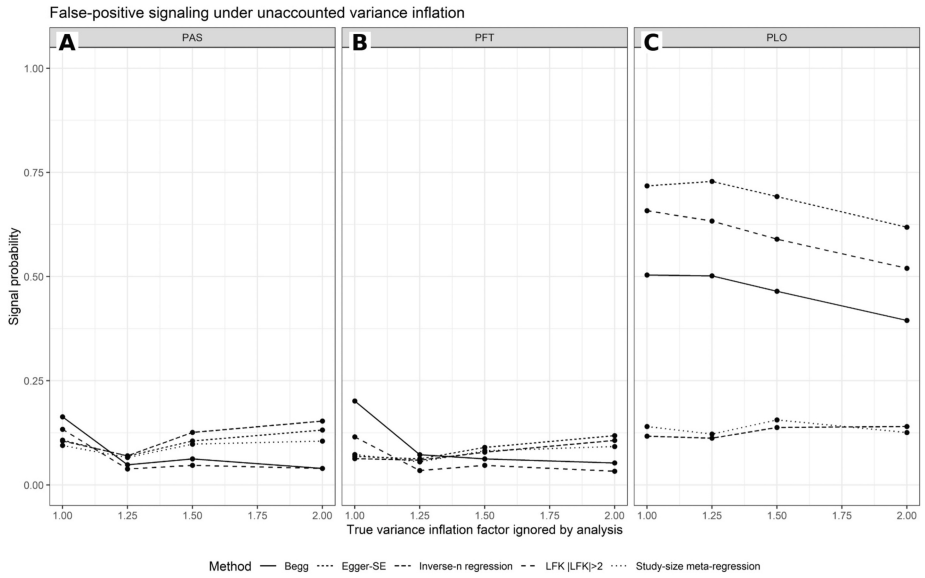


FIGURE 6: Legacy sensitivity to unaccounted variance inflation.

Panel A: PAS; B: PFT; C: PLO. Signal rates after normal perturbations increased transformed-effect variance while reported sampling variances were left unchanged; the displayed subset contains neither a generated size association nor selection. This does not explicitly simulate paired or clustered counts. This separate earlier Version 2.3 sensitivity was not part of the variance-isolated primary comparison and did not determine roadmap ratings.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

Baseline prevalence 5-50%: design validation and controlled anchor

The 5-50% extension used the same K, study-size, heterogeneity, SSE, selection, and transformation architecture as the variance-isolated rare platform, with p0 sampled log-uniformly from 0.05 to 0.50. The validation file confirmed an observed p0 range of 0.050002-0.499997, with all 40,000 underlying datasets inside the prespecified bounds. Generic variance inflation was exactly 1.00 in all 120,000 transformation-specific rows. The selection-intercept calibration used a score-adaptive root bracket to avoid numerical failure when high prevalence and large sample size generated extreme event-excess scores; the target-retention objective was unchanged.

In the controlled anchor, after pooling K = 10, 20, and 40 within each fixed τ level, PAS background signal rates ranged from 2.9% to 7.4% across Begg, Egger-SE, inverse-n regression, the legacy LFK threshold, and direct study-size meta-regression; PFT ranged from 2.9% to 7.2% (Figure 7). Under PLO, Egger-SE increased from 13.4% at $\tau = 0$ to 82.1% at $\tau = 1.20$, Begg from 12.6% to 49.6%, and LFK threshold exceedance from 8.3% to 67.4%. Inverse-n regression remained between 4.5% and 6.9%, and direct study-size meta-regression between 4.3% and 7.1%. Trim-and-fill was excluded from these diagnostic ranges because its k0 > 0 output was descriptive and occurred frequently even in the anchor. The PLO artifact therefore persisted when zero-event studies were uncommon.

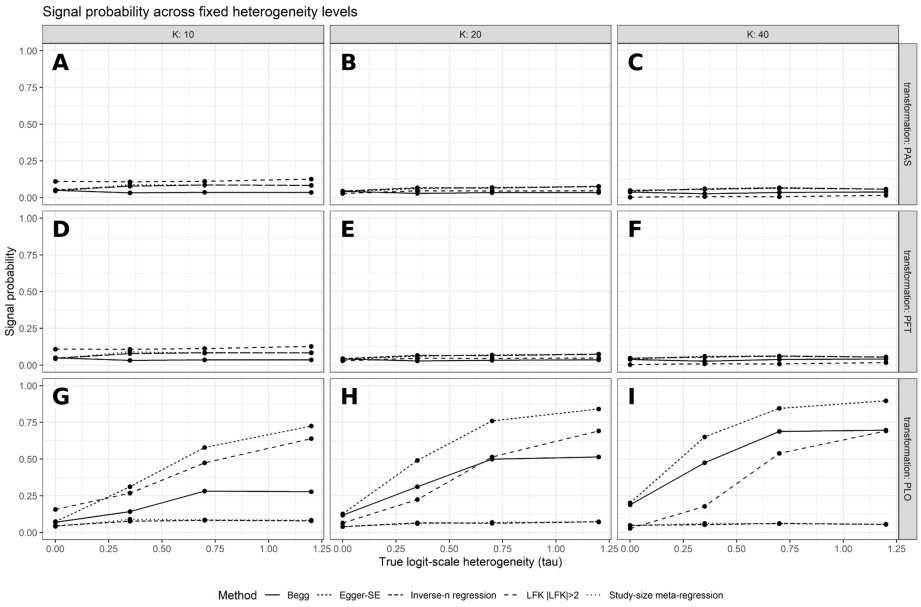


FIGURE 7: Controlled heterogeneity anchor at 5–50% baseline prevalence.

Panels A/B/C show PAS at K = 10/20/40; D/E/F show PFT at K = 10/20/40; G/H/I show PLO at K = 10/20/40. The generic variance-inflation factor was 1.00; neither a genuine SSE nor selection was imposed.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

Figure *7A*-*7C* show PAS at K = 10/20/40, Figure *7D*-*7F* show PFT at those study counts, and Figure *7G*-*7I* show PLO. Comparison of the rows confirms that the PLO-related operating pattern persists outside rare-event settings, while study-size-based regressions remain comparatively stable.

Baseline prevalence 5-50%: data-generating mechanisms

Under PAS, signal rates in the neither and selection-only conditions were, respectively, 2.8% and 4.5% for Begg, 6.1% and 8.0% for Egger-SE, 5.2% and 8.1% for inverse-n regression, 4.0% and 3.8% for LFK, and 5.6% and 8.0% for direct study-size meta-regression (Figure *8*). PFT values were closely similar: 3.0% and 4.7%, 5.6% and 8.0%, 5.4% and 7.7%, 4.2% and 3.7%, and 5.9% and 7.9%. With SSE only, PAS/PFT Egger-SE, inverse-n, and direct study-size meta-regression signaled in 78.5-80.5%, Begg in approximately 69.0%, and LFK in 42.5-42.8%. When SSE and selection coexisted, the corresponding ranges were 75.1-76.9%, approximately 65.7%, and 35.7-36.0%.

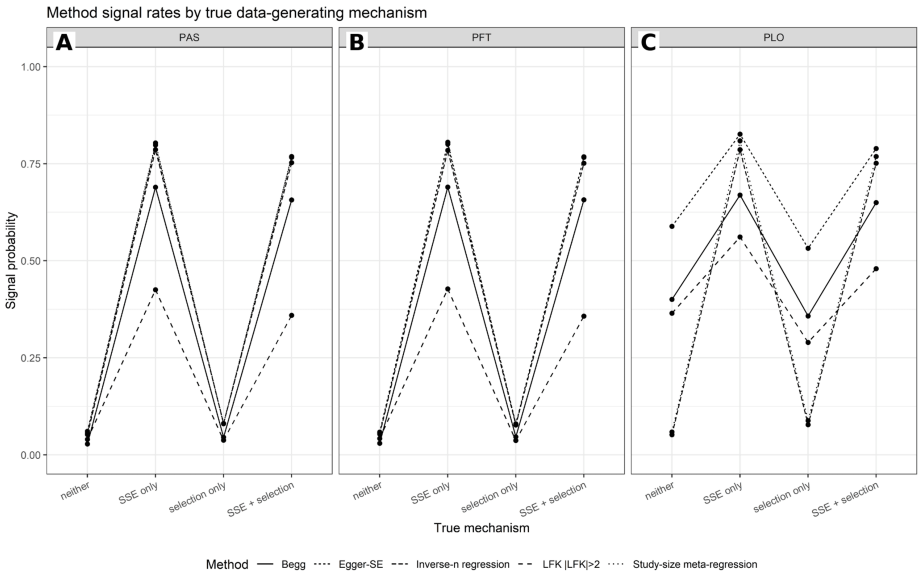


FIGURE 8: Signal rates by data-generating mechanism at 5–50% baseline prevalence.

Panel A: PAS; B: PFT; C: PLO. From left to right, signal probabilities are shown for neither mechanism, SSE only, selection only, and both mechanisms. The same underlying evidence bases were analyzed on all three scales.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

Under PLO, background signaling remained high for effect-SE, rank, and geometry-based procedures. In the neither condition, Egger-SE signaled in 58.9%, Begg in 40.0%, and LFK in 36.5%; under selection only, the corresponding rates were 53.2%, 35.8%, and 29.0%. Direct study-size meta-regression remained at 5.9% and 8.8%, and inverse-n regression at 5.2% and 7.8%, in the same two conditions. With SSE only, PLO signal rates were 82.6% for Egger, 80.9% for direct study-size meta-regression, 78.6% for inverse-n regression, 66.9% for Begg, and 56.1% for LFK. When SSE and selection coexisted, the corresponding values were 78.9%, 76.9%, 75.1%, 65.0%, and 48.0%. The small contrast between selection only and neither under PAS/PFT confirms that these procedures mainly responded to an effect-size-by-study-size pattern rather than to selection as such.

Figure *8A* (PAS), Figure *8B* (PFT), and Figure *8C* (PLO) show the four mechanism states. The first two panels chiefly respond to a generated size-prevalence association, while the PLO panel also shows appreciable Egger, Begg, and LFK signaling when neither mechanism is present.

Baseline prevalence 5-50%: prevalence bands and heterogeneity

Figure *9* stratifies signal probability by baseline-prevalence band and heterogeneity. Under PAS/PFT, the dominant gradient for the formal regression procedures was decreasing signaling as heterogeneity moved from none/low to extreme; differences across the 5-10%, 10-20%, 20-30%, 30-40%, and 40-50% prevalence bands were comparatively modest. LFK was less responsive overall under PAS/PFT and declined with heterogeneity. Begg showed a similar broad heterogeneity gradient.

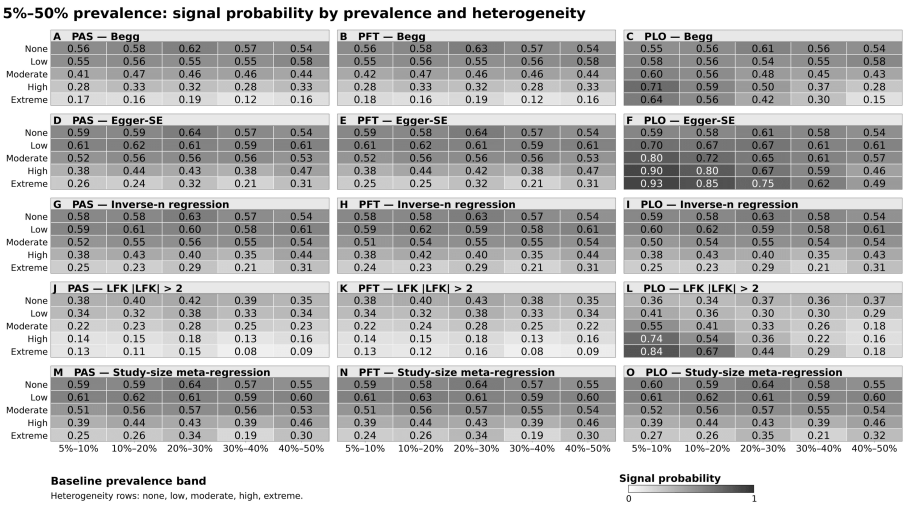


FIGURE 9: Signal probability by baseline-prevalence band and heterogeneity at 5–50% prevalence.

Panels A/B/C: Begg; D/E/F: Egger-SE; G/H/I: inverse-n regression; J/K/L: LFK; M/N/O: direct study-size meta-regression. Within every triplet, panels are PAS, PFT, and PLO, respectively. Five baseline-prevalence bands and five true-τ bands are displayed. Rates are among evaluable stochastic analyses and combine all four mechanism states.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

PLO created a different pattern. Egger-SE and Begg had very high signal probabilities at low prevalence combined with high/extreme heterogeneity and then declined as baseline prevalence approached 50%; nevertheless, their rates remained substantially above the PAS/PFT background. LFK showed the same vulnerability, especially in the lower prevalence bands. Inverse-n and study-size meta-regression were nearly identical across PAS/PFT/PLO within each prevalence-by-heterogeneity cell, supporting their relative transformation stability.

Figure *9A*-*9C* show Begg, D/E/F Egger-SE, G/H/I inverse-n regression, J/K/L LFK, and M/N/O direct study-size meta-regression. PAS, PFT, and PLO occupy the first, second, and third columns, respectively. Differences across prevalence bands are modest for the size-based regressions, whereas the lower-prevalence/high-heterogeneity PLO cells are particularly reactive for effect-precision, rank, and geometric procedures.

Baseline prevalence 5-50%: event yield and heterogeneity

The widened event-yield heatmap (Figure *10*) showed that PAS/PFT regression-based signal rates generally increased from the very low event-yield strata toward moderate/high event yield when heterogeneity was none to moderate. Under high or extreme heterogeneity, signal probabilities were attenuated even when event yield was large. Because the heatmap combines all data-generating mechanisms, it should be interpreted as an operating surface rather than as a pure power curve.

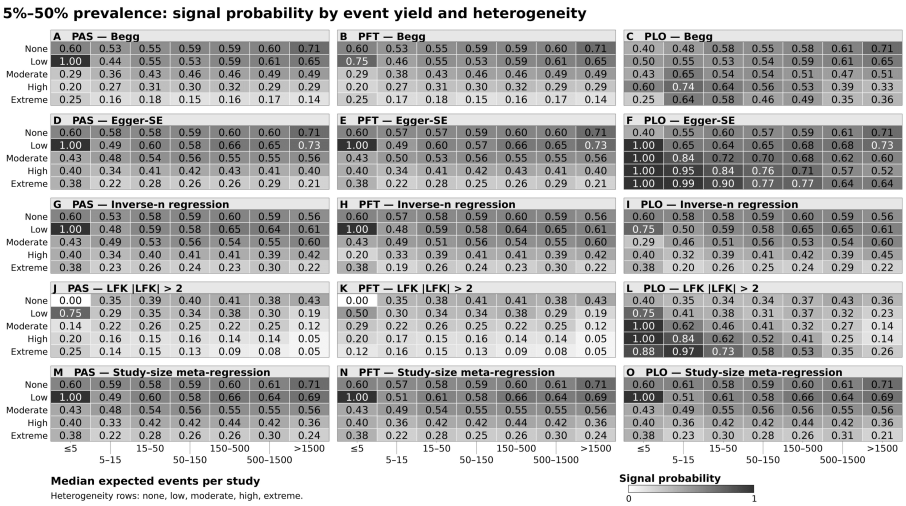


FIGURE 10: Signal probability by event yield and heterogeneity at 5–50% prevalence.

Panels A/B/C: Begg; D/E/F: Egger-SE; G/H/I: inverse-n regression; J/K/L: LFK; M/N/O: direct study-size meta-regression. Within every triplet, panels are PAS, PFT, and PLO, respectively. Expected-event strata are widened for the broader-prevalence domain. Rows run from none to extreme true heterogeneity. The display combines all four mechanisms and is an operating surface, not a pure power curve.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

Under PLO, Egger-SE and Begg were highly reactive in the smallest event-yield strata with high/extreme heterogeneity, while LFK also reached high signal probabilities across several low/intermediate yield cells. Inverse-n and study-size meta-regression retained the same broad event-yield gradient across transformations and did not show the marked PLO inflation seen for effect-SE-based or geometry-based procedures.

Figure *10* uses the same panel allocation: A/B/C Begg, D/E/F Egger-SE, G/H/I inverse-n regression, J/K/L LFK, and M/N/O direct study-size meta-regression, with PAS/PFT/PLO ordered left to right. The widened event-yield strata show attenuation of regression-based signaling with high heterogeneity even when expected event counts are large.

Baseline prevalence 5-50%: correct-direction detection and disagreement

Directional detection by direct study-size meta-regression was substantially higher at 5-50% prevalence than in the rare platform when heterogeneity was absent or low (Figure *11*). At $|\beta_{SSE}| = 0.25$, correct-direction detection across PAS, PFT, and PLO was 98.2-98.6% with no heterogeneity, 91.6-92.5% under low, 66.1-66.7% under moderate, 37.2-37.6% under high, and 18.4-19.6% under extreme heterogeneity. At $|\beta_{SSE}| = 0.75$, detection was 100% with no heterogeneity, 99.7% under low, 96.9-97.4% under moderate, 86.9-87.6% under high, and 69.7-70.6% under extreme heterogeneity. The three transformation curves were nearly superimposable.

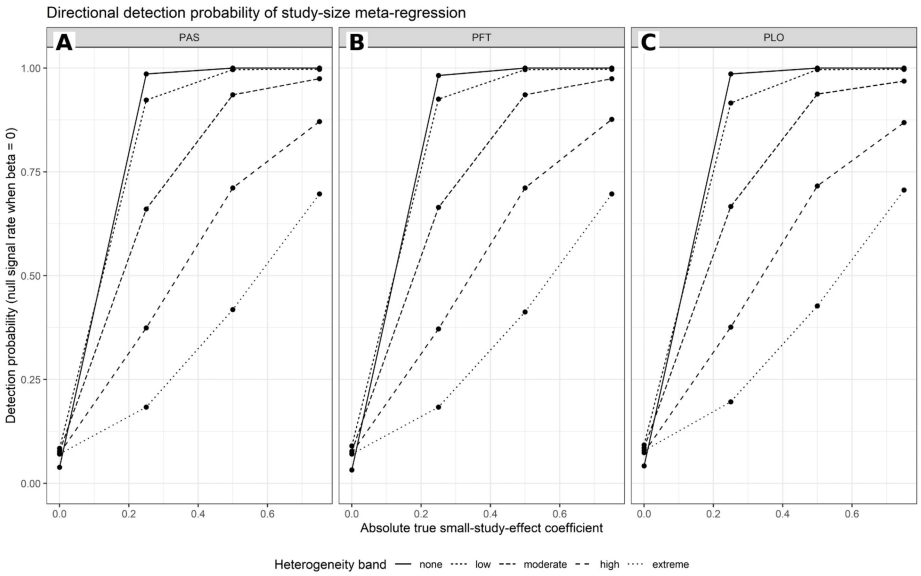


FIGURE 11: Correct-direction detection by study-size meta-regression at 5–50% prevalence.

Panel A: PAS; B: PFT; C: PLO. Detection is plotted against absolute generated SSE coefficient and true heterogeneity. At a generated coefficient of zero, rates represent two-sided background signaling.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

Figure *11A* (PAS), Figure *11B* (PFT), and Figure *11C* (PLO) show nearly superimposable correct-direction detection curves. The study-size-based association is readily detected under none/low heterogeneity, but increasing heterogeneity reduces detection even in this higher-prevalence domain.

Transformation classification disagreement was concentrated in comparisons involving PLO (Figure *12*). PAS-versus-PFT disagreement was 0.57% for Begg, 0.61% for Egger-SE, 0.94% for inverse-n regression, 0.67% for LFK, and 0.67% for direct study-size meta-regression; trim-and-fill disagreement was 2.57%. PAS-versus-PLO and PFT-versus-PLO disagreement was 33.0-33.1% for Begg, 37.9-38.2% for Egger, 38.0-38.3% for LFK, and 27.69-27.73% for trim-and-fill. The corresponding ranges were only 1.6-1.9% for inverse-n and 1.5-2.0% for direct study-size meta-regression. Any disagreement across the three transformations occurred in 38.46% for LFK, 38.38% for Egger, 33.35% for Begg, 28.99% for trim-and-fill, 2.24% for inverse-n, and 2.08% for direct study-size meta-regression. These results support sensitivity-only ratings for PLO-based Egger and Begg. PLO-based LFK is also sensitivity-only except in the 5-50% none/low-heterogeneity stratum, where it is an adjunct.

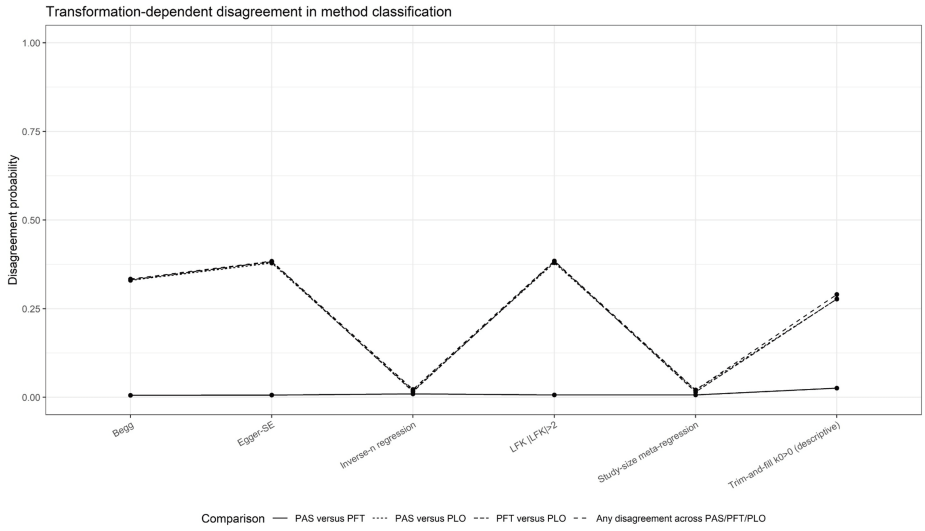


FIGURE 12: Transformation-dependent disagreement at 5–50% prevalence.

Probability that the same dataset receives different binary classifications under PAS, PFT, and PLO. Pairwise disagreements and any disagreement across all three transformations are shown for each method. Trim-and-fill classifications are descriptive.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit; SSE, generated size–prevalence association

True logit-scale heterogeneity and observed I^2

The controlled-anchor crosswalk demonstrated a strong dependence of estimated I^2 on prevalence domain, transformation, and study count, even though the same true logit-scale τ values were generated. With true $\tau = 0$, median I^2 across transformation-by-K cells ranged from 0.0% to 11.6% in the rare/low-prevalence platform and from 0.0% to 1.0% in the 5-50% platform; nevertheless, cell-specific 95th percentiles reached 58.4% and 47.9%. At true $\tau = 0.35$, rare/low-prevalence medians ranged from 9.9% to 46.8%, whereas all broader-prevalence medians were approximately 91% (90.8-91.3%). At $\tau = 0.70$ and 1.20, medians were 47.4-71.5% and 80.2-89.8% in the rare platform, compared with 97.3-97.6% and 99.0-99.2% in the broader platform (Tables *1*, *2*; Figure *13*).

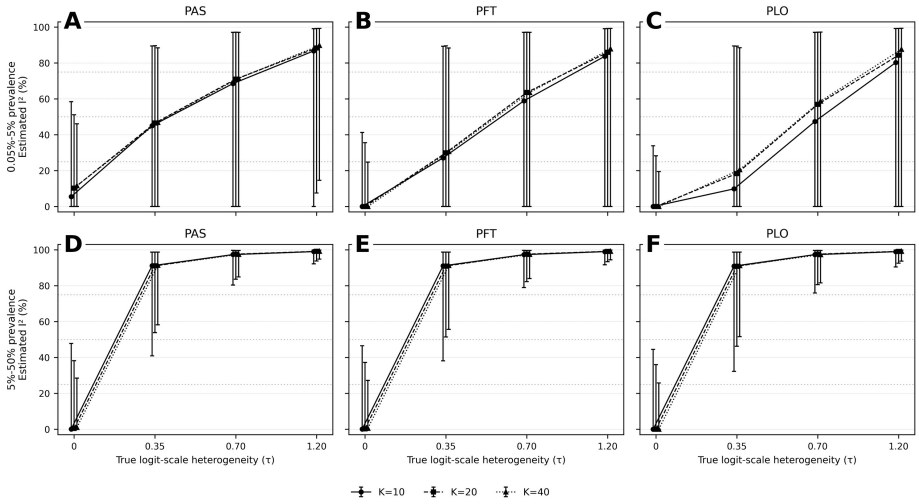


FIGURE 13: Empirical relationship between true logit-scale heterogeneity and estimated I².

Panels A/B/C show rare/low-prevalence PAS/PFT/PLO; D/E/F show 5–50% PAS/PFT/PLO. Points are medians and bars are 5th–95th percentiles across 2,500 replicates per K-by-τ cell; marker and line type identify K = 10, 20, or 40. Horizontal 25%, 50%, and 75% lines are descriptive only. This is an orientation crosswalk, not a conversion chart.

Domain	Scale	True τ	K	Median I² (%)	5th percentile I² (%)	95th percentile I² (%)
0.05–5%	PAS	0.00	10	5.5	0.0	58.4
0.05–5%	PAS	0.00	20	10.3	0.0	51.2
0.05–5%	PAS	0.00	40	11.6	0.0	46.2
0.05–5%	PAS	0.35	10	45.0	0.0	89.5
0.05–5%	PAS	0.35	20	46.5	0.0	89.7
0.05–5%	PAS	0.35	40	46.8	0.0	88.5
0.05–5%	PAS	0.70	10	68.5	0.0	97.1
0.05–5%	PAS	0.70	20	71.0	0.0	97.2
0.05–5%	PAS	0.70	40	71.5	0.0	97.1
0.05–5%	PAS	1.20	10	86.8	0.0	99.1
0.05–5%	PAS	1.20	20	88.3	7.6	99.2
0.05–5%	PAS	1.20	40	89.8	14.6	99.3
0.05–5%	PFT	0.00	10	0.0	0.0	41.3
0.05–5%	PFT	0.00	20	0.0	0.0	35.6
0.05–5%	PFT	0.00	40	0.0	0.0	24.8
0.05–5%	PFT	0.35	10	27.1	0.0	89.3
0.05–5%	PFT	0.35	20	30.0	0.0	89.6
0.05–5%	PFT	0.35	40	30.6	0.0	88.4
0.05–5%	PFT	0.70	10	58.9	0.0	97.1
0.05–5%	PFT	0.70	20	63.6	0.0	97.2
0.05–5%	PFT	0.70	40	63.6	0.0	97.1

0.05–5%	PFT	1.20	10	83.7	0.0	99.1
0.05–5%	PFT	1.20	20	85.9	0.0	99.2
0.05–5%	PFT	1.20	40	87.9	0.0	99.3
0.05–5%	PLO	0.00	10	0.0	0.0	33.9
0.05–5%	PLO	0.00	20	0.0	0.0	28.3
0.05–5%	PLO	0.00	40	0.0	0.0	19.5
0.05–5%	PLO	0.35	10	9.9	0.0	89.5
0.05–5%	PLO	0.35	20	18.4	0.0	89.5
0.05–5%	PLO	0.35	40	20.5	0.0	88.5
0.05–5%	PLO	0.70	10	47.4	0.0	97.1
0.05–5%	PLO	0.70	20	56.9	0.0	97.2
0.05–5%	PLO	0.70	40	58.7	0.0	97.3
0.05–5%	PLO	1.20	10	80.2	0.0	99.2
0.05–5%	PLO	1.20	20	84.4	0.0	99.3
0.05–5%	PLO	1.20	40	87.6	0.0	99.3
5–50%	PAS	0.00	10	0.1	0.0	47.9
5–50%	PAS	0.00	20	0.4	0.0	38.2
5–50%	PAS	0.00	40	1.0	0.0	28.5
5–50%	PAS	0.35	10	91.2	40.9	98.8
5–50%	PAS	0.35	20	91.0	53.8	98.8
5–50%	PAS	0.35	40	91.3	58.2	98.7
5–50%	PAS	0.70	10	97.4	80.4	99.7
5–50%	PAS	0.70	20	97.6	83.6	99.7
5–50%	PAS	0.70	40	97.5	84.9	99.7
5–50%	PAS	1.20	10	99.0	92.2	99.9
5–50%	PAS	1.20	20	99.2	93.8	99.9
5–50%	PAS	1.20	40	99.2	94.8	99.9
5–50%	PFT	0.00	10	0.1	0.0	46.6
5–50%	PFT	0.00	20	0.3	0.0	37.2
5–50%	PFT	0.00	40	0.5	0.0	27.2
5–50%	PFT	0.35	10	91.1	38.1	98.8
5–50%	PFT	0.35	20	91.0	51.4	98.8
5–50%	PFT	0.35	40	91.3	55.7	98.7
5–50%	PFT	0.70	10	97.3	79.0	99.7
5–50%	PFT	0.70	20	97.6	82.3	99.7
5–50%	PFT	0.70	40	97.4	84.0	99.7
5–50%	PFT	1.20	10	99.0	91.6	99.9
5–50%	PFT	1.20	20	99.2	93.4	99.9
5–50%	PFT	1.20	40	99.2	94.5	99.9
5–50%	PLO	0.00	10	0.0	0.0	44.5

5–50%	PLO	0.00	20	0.0	0.0	36.0
5–50%	PLO	0.00	40	0.0	0.0	25.8
5–50%	PLO	0.35	10	91.0	32.2	98.7
5–50%	PLO	0.35	20	90.8	46.3	98.8
5–50%	PLO	0.35	40	91.3	51.6	98.7
5–50%	PLO	0.70	10	97.3	76.0	99.7
5–50%	PLO	0.70	20	97.6	80.6	99.7
5–50%	PLO	0.70	40	97.4	81.6	99.7
5–50%	PLO	1.20	10	99.0	90.5	99.9
5–50%	PLO	1.20	20	99.1	92.6	99.9
5–50%	PLO	1.20	40	99.2	93.7	99.9

TABLE 1: Empirical crosswalk between true logit-scale τ and estimated I^2: study-count-specific summaries.

Each K-by-τ cell comprised 2,500 underlying meta-analyses. Values summarize models with finite I^2; cell evaluability ranged from 90.4% to 100.0%. The 5th and 95th percentiles describe the simulation distribution, not a confidence interval for its median. These values are descriptive and must not be used to convert empirical I^2 into a roadmap category.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit

Domain	Scale	True τ	I² <25% (%)	I² 25–<50% (%)	I² 50–<75% (%)	I² ≥75% (%)
0.05–5%	PAS	0.00	68.9	24.7	6.2	0.1
0.05–5%	PAS	0.35	27.9	27.1	29.0	16.0
0.05–5%	PAS	0.70	16.1	13.3	26.5	44.2
0.05–5%	PAS	1.20	8.4	6.4	15.5	69.8
0.05–5%	PFT	0.00	89.6	9.2	1.1	0.0
0.05–5%	PFT	0.35	46.7	19.6	18.8	14.9
0.05–5%	PFT	0.70	28.3	13.4	19.8	38.6
0.05–5%	PFT	1.20	14.9	7.7	14.1	63.3
0.05–5%	PLO	0.00	94.0	5.6	0.4	0.0
0.05–5%	PLO	0.35	55.5	14.8	15.7	14.0
0.05–5%	PLO	0.70	34.4	12.7	17.3	35.6
0.05–5%	PLO	1.20	15.7	8.2	15.1	61.0
5–50%	PAS	0.00	86.2	12.1	1.7	0.0
5–50%	PAS	0.35	1.2	3.3	15.2	80.3
5–50%	PAS	0.70	0.0	0.1	1.6	98.3
5–50%	PAS	1.20	0.0	0.0	0.1	99.9
5–50%	PFT	0.00	86.9	11.6	1.5	0.0
5–50%	PFT	0.35	1.5	3.5	15.3	79.6
5–50%	PFT	0.70	0.1	0.2	2.0	97.7
5–50%	PFT	1.20	0.0	0.1	0.1	99.8
5–50%	PLO	0.00	88.0	10.6	1.4	0.0
5–50%	PLO	0.35	1.9	4.7	15.1	78.4
5–50%	PLO	0.70	0.1	0.3	2.8	96.8
5–50%	PLO	1.20	0.1	0.0	0.2	99.7

TABLE 2: Descriptive I²-band distributions pooled over study counts in the controlled anchors.

Within each domain, scale, and true-τ stratum, percentages pool K = 10, 20, and 40 among models with finite I². They are not K-specific percentages. Rows may not sum to 100% because of rounding. These descriptive bands were not used to derive or change the roadmap ratings.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit

The stochastic $\tau = 0$ analyses showed why even the anchor crosswalk cannot be used as a universal conversion. When neither a genuine size-prevalence association nor selection was generated, the median I^2 ranged from 0.0% to 4.1% across transformations in the rare platform and from 0.0% to 0.8% in the broader platform. Selection alone produced only modest increases. By contrast, a genuine size-prevalence association without random between-study dispersion produced median I^2 values of 27.3-52.5% in the rare platform and 93.2-93.7% in the broader platform; corresponding values when the genuine association and selection coexisted were 30.2-50.3% and 93.6-94.1%. An intercept-only random-effects model therefore incorporates systematic unmodeled structure into observed inconsistency.

Accordingly, I^2 was included as an applied orientation measure but did not replace true τ in the simulation design or rating derivation. In an empirical anatomical meta-analysis, the practical heterogeneity profile should be interpreted using estimated model-scale τ^2, its uncertainty, the prediction interval, I^2, event yield, study-level estimates, and anatomically plausible moderators. A low I^2 does not prove negligible biological dispersion, and a high I^2 does not by itself identify extreme logit-scale τ or invalidate pooling.

Table 1 presents the study-count-specific medians and percentiles. Table 2 reports the descriptive I^2-band percentages pooled over K = 10, 20, and 40; those pooled percentages must not be interpreted as K-specific estimates.

Figure 13A, 13B, and 13C show the rare/low-prevalence PAS, PFT, and PLO crosswalks, respectively; Figure 13D, 13E, and 13F show the corresponding 5-50% crosswalks. The contrast between the upper and lower rows demonstrates why empirical I^2 cannot be converted mechanically into a true-τ roadmap stratum.

One-stage BN-GLMM: computational integrity and fit health

All 80,000 underlying evidence bases were regenerated without material mismatch, and the three primary BN-GLMM specifications yielded 240,000 one-stage fitting runs. All integrated logistic signal-driver models converged without warnings or errors.

For the direct study-size BN-GLMM, evaluability was 98.95% in the rare/low-prevalence anchor and 99.72% in the corresponding stochastic module. Clean-fit rates were 98.16% and 98.87%, and warning rates were 1.14% and 0.94%, respectively. In the 5-50% domain, the direct model was estimable and clean in every anchor and stochastic evidence base, without fitting warnings. All-zero evidence bases occurred in 0.69% of the rare anchor and 0.19% of the rare stochastic module but did not occur in the broader-prevalence platform.

The rare-anchor information-state analysis included 208 all-zero evidence bases, 454 evidence bases with one event-bearing study, 644 with two event-bearing studies, and 28,694 with at least three event-bearing studies. All-zero evidence bases were structurally non-estimable. With one event-bearing study, evaluability was 77.1%, the model-error rate was 22.9%, and the warning rate was 24.7%. The median absolute study-size coefficient was 0.758, its 95th percentile was 4.72, and its maximum was 727.6. With two event-bearing studies, evaluability was 99.7%, but the median, 95th percentile, and maximum absolute coefficients remained 0.671, 2.81, and 41.2, respectively. All evidence bases with at least three event-bearing studies were estimable; their corresponding coefficient summaries were 0.132, 0.686, and 6.55.

These results demonstrate a graded sparse-information problem rather than a universal event-bearing-study threshold. In particular, the observation that all evidence bases with at least three event-bearing studies were estimable should not be converted into a post hoc eligibility rule. It instead indicates that K, total event yield, the number of event-bearing studies, and model warnings should be reported together when interpreting one-stage moderator analyses (Tables 3, 4).

Domain	Module	All-zero (%)	Evaluable (%)	Clean fit (%)	Warning (%)	Median event-bearing studies	5th percentile event-bearing studies
0.05–5%	Controlled anchor	0.69	98.95	98.16	1.14	15	3
0.05–5%	Stochastic module	0.19	99.72	98.87	0.94	19	5
5–50%	Controlled anchor	0.00	100.00	100.00	0.00	20	10
5–50%	Stochastic module	0.00	100.00	100.00	0.00	25	11

TABLE 3: Evaluability and fit health of the one-stage direct study-size BN-GLMM.

Notes: A clean fit was estimable without a standard-error flag or captured warning. All-zero evidence bases were structurally non-estimable and were not counted as model errors. Information-state categories describe the observed information gradient and should not be interpreted as validated universal eligibility thresholds.

BN-GLMM: binomial-normal mixed-effects framework

Domain	Information state	Meta-analyses	Evaluable (%)	Clean fit (%)	Warning (%)	Model error (%)	Median |β_{size}|	95th percentile |β_{size}|	Maximum |β_{size}|
0.05–5%	All zero	208	0.00	0.00	0.00	0.00	NE	NE	NE
0.05–5%	One event-bearing study	454	77.09	75.33	24.67	22.91	0.758	4.720	727.627
0.05–5%	Two event-bearing studies	644	99.69	98.29	1.71	0.31	0.671	2.808	41.200
0.05–5%	Three or more event-bearing studies	28694	100.00	99.23	0.77	0.00	0.132	0.686	6.546
5–50%	Three or more event-bearing studies	30000	100.00	100.00	0.00	0.00	0.069	0.440	2.404

TABLE 4: Sparse-information states and direct BN-GLMM coefficient behavior in the controlled anchors.

Notes: A clean fit was estimable without a standard-error flag or captured warning. All-zero evidence bases were structurally non-estimable and were not counted as model errors. Information-state categories describe the observed information gradient and should not be interpreted as validated universal eligibility thresholds. NE, not estimable; these entries are not zero. Coefficient summaries refer to the absolute direct study-size coefficient.

BN-GLMM: binomial-normal mixed-effects framework

Table *3* summarizes fit health by prevalence domain and module. Table *4* separates sparse-information states, computational outcomes, and coefficient summaries; NE identifies structurally non-estimable coefficient summaries.

One-stage BN-GLMM: calibration, coverage, and directional detection

When true $\tau = 0$, Wald-z signal rates for the direct BN study-size moderator ranged from 3.8% to 4.7% across K = 10, 20, and 40 in the rare domain and from 4.2% to 4.9% in the 5-50% domain. With nonzero heterogeneity, signaling increased most clearly when K = 10. In rare data, rates were 7.9%, 9.1%, and 10.1% at true $\tau = 0.35$, 0.70, and 1.20, respectively. Corresponding broader-domain rates were 11.0%, 11.8%, and 11.4%. At K = 40, the rates were mostly 5.5-6.6%. Thus, the one-stage model avoided the extreme background signaling observed for several continuity-corrected two-stage PLO procedures, but Wald-z moderator inference was mildly anti-conservative when the number of studies was small and heterogeneity was present.

The derived t-reference sensitivity improved small-K calibration. At K = 10 in the rare domain, signal rates at $\tau = 0.35$, 0.70, and 1.20 decreased from 7.9%, 9.1%, and 10.1% to 4.2%, 5.0%, and 6.0%. Corresponding 95% coverage increased from 92.1%, 90.9%, and 89.9% to 95.8%, 95.0%, and 94.0%. In the broader domain, signaling decreased from 11.0%, 11.8%, and 11.4% to 7.2%, 7.1%, and 7.1%, while coverage increased from 89.0%, 88.2%, and 88.6% to 92.8%, 92.9%, and 92.9%. At true $\tau = 0$, however, the t reference was conservative at K = 10, with signaling of approximately 1.4-1.9%. Thus, the improvement did not consistently restore nominal calibration: broader-prevalence settings retained approximately 7% background signaling and 93% coverage. The t reference is an inferential sensitivity, not a generally validated small-sample correction.

Without selective dissemination, correct-direction detection for $|\beta_{SSE}| = 0.25$ across none, low, moderate, high, and extreme heterogeneity was 52.8%, 53.1%, 41.7%, 35.0%, and 12.4% in rare data. Corresponding values in the broader domain were 98.2%, 97.5%, 72.5%, 36.2%, and 24.1%. The derived t-reference sensitivity generally reduced detection by approximately 1-4 percentage points across the aggregated contexts. Table *5* reports K = 10 calibration and coverage, and Table *6* preserves detection results separately for simulations without selection and results weighted across selection states. These are different analysis subsets and should not be compared as if their denominators were identical.

Domain	True τ	Evaluable R	Wald-z signal (%)	t-reference signal (%)	Wald-z coverage (%)	t-reference coverage (%)
0.05–5%	0.00	2430	3.8	1.4	96.2	98.6
0.05–5%	0.35	2423	7.9	4.2	92.1	95.8
0.05–5%	0.70	2445	9.1	5.0	90.9	95.0
0.05–5%	1.20	2454	10.1	6.0	89.9	94.0
5–50%	0.00	2500	4.9	1.9	95.1	98.1
5–50%	0.35	2500	11.0	7.2	89.0	92.8
5–50%	0.70	2500	11.8	7.1	88.2	92.9
5–50%	1.20	2500	11.4	7.1	88.6	92.9

TABLE 5: Background signaling and 95% coefficient coverage for the direct BN-GLMM moderator at K = 10.

The t reference used K − 2 degrees of freedom and was calculated from stored coefficient estimates and standard errors without refitting. A signal was a two-sided p value below 0.05. The controlled anchors generated neither a size–prevalence association nor selective dissemination.

BN-GLMM: binomial-normal mixed-effects framework

Domain	Heterogeneity	Selection scope	$\lvert\beta_{SSE}\rvert$	Evaluable R	Wald-z detection (%)	t-reference detection (%)	Change (percentage points)
0.05–5%	None	No selection	0.25	108	52.8	NR	NR
0.05–5%	None	No selection	0.50	144	79.9	NR	NR
0.05–5%	None	No selection	0.75	74	82.4	NR	NR
0.05–5%	Low	No selection	0.25	160	53.1	NR	NR
0.05–5%	Low	No selection	0.50	205	77.1	NR	NR
0.05–5%	Low	No selection	0.75	109	91.7	NR	NR
0.05–5%	Moderate	No selection	0.25	180	41.7	NR	NR
0.05–5%	Moderate	No selection	0.50	236	69.5	NR	NR
0.05–5%	Moderate	No selection	0.75	151	87.4	NR	NR
0.05–5%	High	No selection	0.25	163	35.0	NR	NR
0.05–5%	High	No selection	0.50	250	65.6	NR	NR
0.05–5%	High	No selection	0.75	147	79.6	NR	NR
0.05–5%	Extreme	No selection	0.25	137	12.4	NR	NR

0.05–5%	Extreme	No selection	0.50	135	37.8	NR	NR
0.05–5%	Extreme	No selection	0.75	74	62.2	NR	NR
5–50%	None	No selection	0.25	109	98.2	NR	NR
5–50%	None	No selection	0.50	144	100.0	NR	NR
5–50%	None	No selection	0.75	75	100.0	NR	NR
5–50%	Low	No selection	0.25	161	97.5	NR	NR
5–50%	Low	No selection	0.50	205	99.5	NR	NR
5–50%	Low	No selection	0.75	109	100.0	NR	NR
5–50%	Moderate	No selection	0.25	182	72.5	NR	NR
5–50%	Moderate	No selection	0.50	236	96.2	NR	NR
5–50%	Moderate	No selection	0.75	152	98.7	NR	NR
5–50%	High	No selection	0.25	163	36.2	NR	NR
5–50%	High	No selection	0.50	251	81.3	NR	NR
5–50%	High	No selection	0.75	147	94.6	NR	NR
5–50%	Extreme	No selection	0.25	137	24.1	NR	NR
5–50%	Extreme	No selection	0.50	135	55.6	NR	NR
5–50%	Extreme	No selection	0.75	74	77.0	NR	NR
0.05–5%	None/low	All states	0.25	688	56.0	53.2	-2.8
0.05–5%	Moderate	All states	0.25	490	43.3	39.0	-4.3
0.05–5%	High/extreme	All states	0.25	765	24.6	22.4	-2.2
5–50%	None/low	All states	0.25	694	95.5	94.2	-1.3
5–50%	Moderate	All states	0.25	492	68.7	66.5	-2.2
5–50%	High/extreme	All states	0.25	765	32.3	29.4	-2.9

TABLE 6: Correct-direction detection by the direct BN-GLMM moderator and derived t-reference sensitivity.

Correct-direction detection required $p < 0.05$ and agreement of the estimated and generated coefficient signs. No-selection rows retain the effect-specific Wald-z results from the original analysis. All-states rows summarize the derived t-reference sensitivity for $|\beta_{SSE}| = 0.25$, weighted across selection states. Change is t-reference minus Wald-z detection. NR, not reported for that subset in this table; it does not mean zero or non-estimability. All-states and no-selection denominators differ by design.

BN-GLMM: binomial-normal mixed-effects framework

Figure *14* displays the background-signal and correct-direction detection summaries for both prevalence domains. The vertical 5% reference applies only to background-signal rows. The one-stage direct moderator predominantly has a power limitation as heterogeneity increases, together with mildly anti-conservative small-K Wald-z inference.

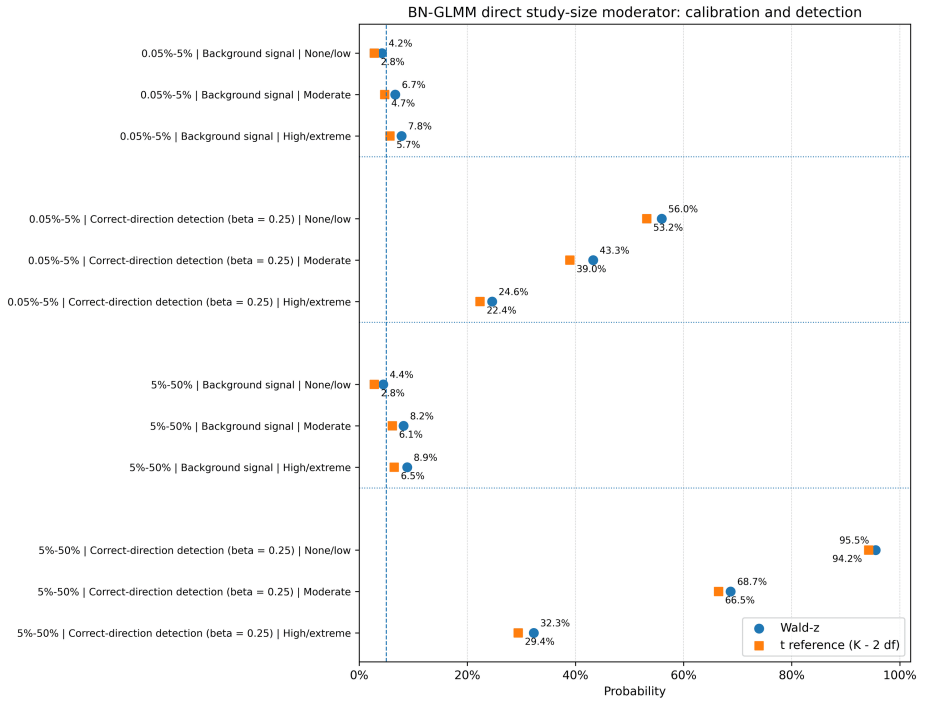


FIGURE 14: Wald-z and derived t-reference operating behavior of the direct BN-GLMM moderator.

Background signal probabilities are pooled over study count within the displayed heterogeneity contexts. Correct-direction probabilities for |βSSE| = 0.25 are weighted across selection states. The vertical 5% line applies only to background-signal rows. The t reference used K − 2 degrees of freedom and was computed from stored estimates and standard errors without refitting.

BN-GLMM: binomial-normal mixed-effects framework

One-stage BN-GLMM: selection, generated size effects, and signal drivers

Weighted across heterogeneity bands, direct BN study-size moderator signal rates under neither mechanism, selection only, a generated size-prevalence association only, and both mechanisms were 5.8%, 8.8%, 61.8%, and 63.5% in rare data. Corresponding rates in the 5-50% domain were 7.2%, 9.4%, 82.1%, and 78.5%. The inverse-n BN moderator produced the same qualitative pattern, with rates of 5.5%, 7.6%, 59.4%, and 60.1% in rare data and 6.4%, 9.8%, 80.1%, and 76.8% in the broader domain.

The adjusted logistic signal-driver models reinforced this distinction. For the direct BN study-size moderator, the presence of a generated size-prevalence association was strongly associated with signaling in rare data (odds ratio (OR), 30.28; 95% CI, 26.28-34.89) and in the broader domain (OR, 82.92; 95% CI, 71.09-96.72). Selective dissemination was not independently associated with signaling in either domain (rare: OR, 1.08; 95% CI, 0.97-1.20; broader: OR, 0.90; 95% CI, 0.80-1.02). The OR per one-standard-deviation increase in true τ was 0.61 (95% CI, 0.58-0.64) in rare data and 0.39 (95% CI, 0.36-0.41) in broader data, consistent with reduced power as heterogeneity increased. The inverse-n moderator showed the same substantive pattern (Table 7).

Domain	Moderator	Predictor	Adjusted OR	95% CI lower	95% CI upper	p-value
0.05–5%	Direct study-size	Generated size-prevalence association present	30.28	26.28	34.89	<0.001
0.05–5%	Direct study-size	Selective dissemination present	1.08	0.97	1.20	0.138
0.05–5%	Direct study-size	True logit-scale τ (per 1 SD increase)	0.61	0.58	0.64	<0.001
0.05–5%	Inverse-n	Generated size-prevalence association present	30.29	26.15	35.08	<0.001
0.05–5%	Inverse-n	Selective dissemination present	1.01	0.91	1.12	0.872
0.05–5%	Inverse-n	True logit-scale τ (per 1 SD increase)	0.60	0.56	0.63	<0.001
5–50%	Direct study-size	Generated size-prevalence association present	82.92	71.09	96.72	<0.001
5–50%	Direct study-size	Selective dissemination present	0.90	0.80	1.02	0.089
5–50%	Direct study-size	True logit-scale τ (per 1 SD increase)	0.39	0.36	0.41	<0.001
5–50%	Inverse-n	Generated size-prevalence association present	71.05	61.15	82.56	<0.001
5–50%	Inverse-n	Selective dissemination present	0.95	0.84	1.07	0.384
5–50%	Inverse-n	True logit-scale τ (per 1 SD increase)	0.39	0.36	0.41	<0.001

TABLE 7: Adjusted associations of generated evidence characteristics with BN-GLMM moderator signaling.

Notes: Odds ratios were obtained from method-specific logistic models. The printed table shows the three principal explanatory terms. Every model additionally adjusted for the natural logarithm of the number of studies, the natural logarithm of one plus the median expected number of events per study, and standardized within-meta-analysis variability in log-transformed study size. The odds ratio for true heterogeneity represents a one-standard-deviation increase. A signal denotes fulfillment of the method-specific criterion and does not diagnose publication bias.

BN-GLMM: binomial-normal mixed-effects framework

These findings indicate that the one-stage moderator principally detected the generated size-prevalence relationship rather than selective dissemination itself. Selection could therefore alter the available prevalence evidence without necessarily creating the model-based size pattern required for a positive moderator result.

One-stage versus two-stage agreement and heterogeneity recovery

Classification agreement between the direct BN study-size moderator and the PAS, PFT, and PLO direct study-size regressions was 93.3%, 96.0%, and 91.1% in the rare anchor and 90.7%, 94.1%, and 89.4% in the rare stochastic module. Corresponding agreement was 98.1%, 98.2%, and 98.1% in the broader anchor and 98.0%, 98.0%, and 98.2% in its stochastic module.

In the rare anchor, the direct BN moderator signaled in approximately 6.5% of paired evidence bases, compared with 11.3% for continuity-corrected two-stage PLO. In the broader anchor, BN signaling was 7.6%, compared with approximately 6.3-6.4% for the two-stage regressions. The one-stage model was therefore not uniformly less reactive. Its principal advantage was observed when sparse boundary data interacted with the continuity-corrected two-stage PLO implementation.

At 5-50% prevalence, direct BN and two-stage PLO coefficients were nearly identical, with median paired differences of approximately −0.001 to −0.002 and root mean squared differences of approximately 0.02. Coefficient agreement was weaker in rare data because occasional information-poor evidence bases generated extreme estimates. Numerical coefficient comparisons with PAS and PFT were not interpreted because their coefficients were expressed on different transformation scales.

Both logit-scale approaches underestimated true heterogeneity in rare data, particularly with K = 10, but the BN-GLMM was generally less downward-biased. At K = 10 and true τ = 0.35, 0.70, and 1.20, median BN estimates were 0.244, 0.563, and 1.045, compared with 0.152, 0.454, and 0.868 under two-stage PLO. At K = 40, the corresponding medians were 0.317, 0.673, and 1.165 for BN and 0.236, 0.543, and 0.965 for PLO. In the 5-50% domain, the approaches were closely comparable: at K = 10, BN medians were 0.317, 0.642, and 1.090, compared with 0.333, 0.667, and 1.121 for PLO; at K = 40, medians were 0.341, 0.686, and 1.178 versus 0.341, 0.683, and 1.160 (Figure *15*; Tables *8*, *9*).

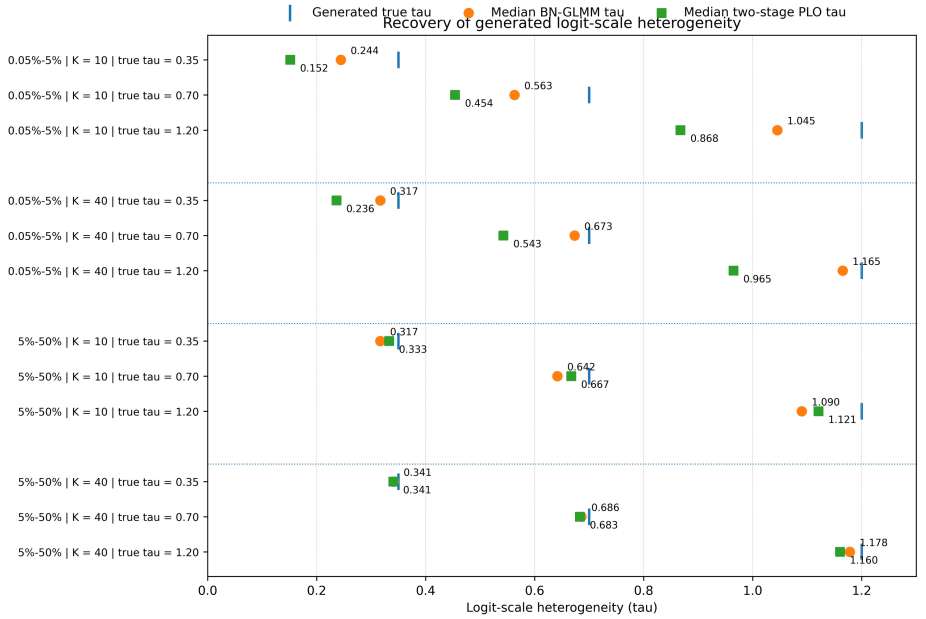


FIGURE 15: Recovery of generated logit-scale heterogeneity by one-stage BN-GLMM and two-stage PLO.

Each row represents a controlled-anchor prevalence-domain, study-count, and nonzero true-τ cell. Vertical ticks identify generated τ; circles and squares show median BN-GLMM and two-stage PLO estimates, respectively. The figure supplements the I^2 crosswalk but does not provide a universal conversion from empirical I^2 to true τ.

BN-GLMM: binomial-normal mixed-effects framework; PLO: continuity-corrected two-stage logit

Domain	Module	Two-stage scale	Agreement (%)	BN signal (%)	Two-stage signal (%)	BN-only (%)	Two-stage-only (%)
0.05–5%	Controlled anchor	PAS	93.3	6.6	8.3	2.5	4.2
0.05–5%	Controlled anchor	PFT	96.0	6.6	6.9	1.9	2.2
0.05–5%	Controlled anchor	PLO	91.1	6.5	11.3	2.1	6.8
0.05–5%	Stochastic module	PAS	90.7	40.3	39.4	5.1	4.2
0.05–5%	Stochastic module	PFT	94.1	40.3	39.3	3.4	2.5
0.05–5%	Stochastic module	PLO	89.4	40.4	41.8	4.6	6.0
5–50%	Controlled anchor	PAS	98.1	7.6	6.4	1.6	0.4
5–50%	Controlled anchor	PFT	98.2	7.6	6.3	1.5	0.2
5–50%	Controlled anchor	PLO	98.1	7.6	6.4	1.5	0.3
5–50%	Stochastic module	PAS	98.0	50.8	49.2	1.8	0.2
5–50%	Stochastic module	PFT	98.0	50.8	49.2	1.8	0.2
5–50%	Stochastic module	PLO	98.2	50.8	49.5	1.5	0.3

TABLE 8: Paired classification agreement between one-stage and two-stage direct study-size analyses.

Comparisons paired methods fitted to identical evidence bases and were restricted to mutually evaluable results. BN-only and two-stage-only values were derived from paired agreement and marginal signal probabilities; small discrepancies between sums are due to rounding. A signal denotes $p < 0.05$, not a diagnosis of publication bias.

PAS, ordinary arcsine-square-root; PFT, Freeman–Tukey double-arcsine; PLO, continuity-corrected two-stage logit

Domain	K	True τ	Median BN τ	Median two-stage PLO τ	Median BN bias	Median PLO bias
0.05–5%	10	0.35	0.244	0.152	-0.106	-0.198
0.05–5%	10	0.70	0.563	0.454	-0.137	-0.246
0.05–5%	10	1.20	1.045	0.868	-0.155	-0.332
0.05–5%	40	0.35	0.317	0.236	-0.033	-0.114
0.05–5%	40	0.70	0.673	0.543	-0.027	-0.157
0.05–5%	40	1.20	1.165	0.965	-0.035	-0.235
5–50%	10	0.35	0.317	0.333	-0.033	-0.017
5–50%	10	0.70	0.642	0.667	-0.058	-0.033
5–50%	10	1.20	1.090	1.121	-0.110	-0.079
5–50%	40	0.35	0.341	0.341	-0.009	-0.009
5–50%	40	0.70	0.686	0.683	-0.014	-0.017
5–50%	40	1.20	1.178	1.160	-0.022	-0.040

TABLE 9: Median logit-scale heterogeneity recovery in selected controlled-anchor cells.

The intercept-only BN-GLMM and two-stage PLO estimate heterogeneity on the common logit scale. Bias is the median estimated τ minus the generated true τ; it is not the mean estimation error. Negative values indicate underestimation. The table reports nonzero true-τ cells at K = 10 and 40.

BN-GLMM, binomial-normal mixed-effects framework; PLO, continuity-corrected two-stage logit

The logit-scale comparison strengthens the interpretation of heterogeneity recovery but does not provide a universal conversion between empirical I^2 and true τ.

Table *8* gives paired classification results on mutually evaluable evidence bases, and Table *9* gives median heterogeneity recovery on the common logit scale. Figure *15* contrasts the median BN-GLMM and two-stage PLO estimates against the generated nonzero τ values. The displayed bias summaries are median estimate minus true τ, not mean bias.

Integrated practical roadmap

Table *10* combines the one-stage BN-GLMM moderator and transformation-specific two-stage ratings in a single matrix. Count-based moderator analysis should be considered first when model assumptions, event information, and fit diagnostics permit; the two-stage rows supply transformation-specific triangulation. The ratings require at least 10 studies and meaningful study-size variation. With fewer than 10 studies, no roadmap cell applies: pooling may still be defensible anatomically and methodologically, but the evidence lies outside the evaluated domain of formal asymmetry inference.

Domain	Framework or scale	True-τ context	Direct study size	Inverse-n	Egger-SE	Begg	Doi/LFK	Trim-and-fill
0.05–5%	BN-GLMM	None/low	P	A	NA	NA	NA	NA
0.05–5%	BN-GLMM	Moderate	A	A	NA	NA	NA	NA
0.05–5%	BN-GLMM	High/extreme	A	A	NA	NA	NA	NA
5–50%	BN-GLMM	None/low	P	A	NA	NA	NA	NA
5–50%	BN-GLMM	Moderate	P	A	NA	NA	NA	NA
5–50%	BN-GLMM	High/extreme	A	A	NA	NA	NA	NA
0.05–5%	PAS	None/low	P	P	P	A	A	S
0.05–5%	PAS	Moderate	A	A	A	A	A	S
0.05–5%	PAS	High/extreme	A	A	A	A	A	S
0.05–5%	PFT	None/low	P	P	P	A	A	S
0.05–5%	PFT	Moderate	A	A	A	A	A	S
0.05–5%	PFT	High/extreme	A	A	A	A	A	S
0.05–5%	PLO	None/low	A	A	S	S	S	S
0.05–5%	PLO	Moderate	A	A	S	S	S	S
0.05–5%	PLO	High/extreme	A	A	S	S	S	S
5–50%	PAS	None/low	P	P	P	A	A	S
5–50%	PAS	Moderate	P	P	P	A	A	S
5–50%	PAS	High/extreme	A	A	A	A	A	S
5–50%	PFT	None/low	P	P	P	A	A	S
5–50%	PFT	Moderate	P	P	P	A	A	S
5–50%	PFT	High/extreme	A	A	A	A	A	S
5–50%	PLO	None/low	P	P	S	S	A	S
5–50%	PLO	Moderate	P	P	S	S	S	S
5–50%	PLO	High/extreme	A	A	S	S	S	S

TABLE 10: Integrated simulation-informed roadmap for one-stage moderator and two-stage asymmetry analyses.

Ratings require K ≥ 10, meaningful size/precision variation and, for BN-GLMM, sufficient events and acceptable fit diagnostics. The intercept-only model is not an asymmetry procedure; BN inverse-n is an adjunct throughout. True-τ contexts cannot be inferred from I^2. PFT/PLO ratings do not endorse primary prevalence estimators. Labels retain original point estimates: near 50% detection, an overlapping Monte Carlo interval warrants adjunct interpretation. P does not guarantee 5% signaling or 95% coverage. No rating establishes publication bias.

P, conditional primary candidate; A, adjunct; S, sensitivity/descriptive only; NA, not applicable; BN-GLMM, binomial-normal mixed-effects framework; PFT, Freeman-Tukey double-arcsine; PLO, continuity-corrected two-stage logit

The most important cross-range finding is that transformation and heterogeneity outweighed baseline prevalence as determinants of asymmetry behavior. Under PAS/PFT, direct study-size, inverse-n, and Egger analyses were preferred with none/low heterogeneity in both prevalence domains. At rare/low prevalence, they were downgraded to adjunct from moderate heterogeneity onward because correct-direction detection of $|\beta_{SSE}| = 0.25$ was only 41.4-43.3% under moderate heterogeneity. At 5-50% prevalence, they remained preferred under moderate heterogeneity, where the corresponding detection was 66.1-66.7%, but were downgraded under high/extreme heterogeneity, where it fell to 18.4-37.6%. Begg and LFK remained adjuncts under PAS/PFT because they supplied supplementary information but were less discriminating or had noninvariant thresholds. Trim-and-fill was sensitivity only in every cell.

The principal prevalence-dependent difference concerned PLO. At rare/low prevalence, even the study-size-based regressions were adjuncts because sparse events and boundary corrections reduced evaluability and stability. At 5-50% prevalence, direct study-size and inverse-n regression were preferred under none/low and moderate heterogeneity and adjunct under high/extreme heterogeneity because their PLO operating curves were nearly identical to PAS/PFT. PLO Egger and Begg remained sensitivity-only because heterogeneity generated high background signals. PLO LFK was adjunct only in the 5-50% none/low-heterogeneity stratum and sensitivity-only thereafter.

For the one-stage direct moderator, rare/low-prevalence none/low-heterogeneity settings are primary-candidate settings with an event-information caution and t-reference sensitivity when K is near 10. Rare moderate heterogeneity warrants adjunct use with explicit inspection of information and warnings, and rare high/extreme heterogeneity carries markedly reduced power. In the 5-50% domain, the direct moderator is a primary candidate under none/low and moderate heterogeneity, subject to the usual model and fit assumptions, and an adjunct under high/extreme heterogeneity. The inverse-n BN moderator is an adjunct throughout because it is an indirect proxy. The observed stability with at least three event-bearing studies is descriptive, not a universal eligibility threshold.

Monte Carlo uncertainty qualifies the boundary-sensitive ratings. For the rare-domain direct BN moderator with $|\beta_{SSE}| = 0.25$, no selection and no heterogeneity, detection was 57/108 (52.8%; MCSE, 4.8 percentage points; 95% Monte Carlo interval, 43.4-61.9%). Across the pooled none/low context and all selection states, Wald-z detection was 385/688 (56.0%; interval, 52.2-59.6%); the t-reference value was 366/688 (53.2%; interval, 49.5-56.9%). These different denominators reflect different subsets. The corresponding all-state PAS direct-regression estimates were 382/687 (55.6%; interval, 51.9-59.3%) for none/low heterogeneity and 203/490 (41.4%; interval, 37.2-45.8%) for moderate heterogeneity. Raising the detection floor to 60% would change the detection component of rare none/low P ratings; lowering it to 40% would change the component for rare moderate settings. These threshold and interval sensitivities favor a graded interpretation, not confidence in a categorical boundary. Original matrix labels are retained as point-estimate summaries, with boundary-sensitive primary-candidate ratings explicitly provisional.

Heterogeneity rows refer to known true τ in the simulations, not to a quantity directly observable in an empirical review. Tables *1*-*2* and Figure *13* show why estimated I^2 cannot mechanically identify a row. Practical interpretation should instead consider model-scale heterogeneity and uncertainty, prediction intervals, event yield, the distribution of study-level prevalence, and anatomically justified sources of heterogeneity. None of the P/A/S labels denotes a stand-alone test of publication bias. When heterogeneity uncertainty spans more than one simulation context, consider the implications of the adjacent contexts and use the more cautious, adjunct interpretation rather than selecting a row from a point estimate alone.

Discussion

Start with the scientific question, not the plot

The first decision is conceptual: does the investigator wish to describe asymmetry, test whether prevalence varies systematically with study size, or assess the risk that relevant evidence is missing? These are different questions. Direct study-size meta-regression addresses the size-prevalence association represented in the simulation. Doi/LFK describes the magnitude and direction of plot asymmetry. Trim-and-fill evaluates how a pooled estimate changes under a particular rank-based imputation algorithm. None of these procedures can determine why studies are absent or establish that unpublished studies exist.

Within the conditions evaluated here, a practical hierarchy is to use an appropriate one-stage count-based random-effects model for primary prevalence synthesis when counts, denominators, and model assumptions permit. Its direct study-size moderator is a conditional primary candidate for a size-prevalence relationship, recognizing that both its functional form and the normal random-effects distribution match the generating model; the corresponding two-stage direct study-size meta-regression provides a transformation-specific comparator. Inverse-sample-size regression and Egger-SE are secondary triangulation, Doi/LFK is a continuous descriptive display, Begg is supplementary, and trim-and-fill is a sensitivity analysis only. Agreement across procedures strengthens evidence that a size-related pattern is present, but it does not identify its cause.

Transformation choice precedes asymmetry assessment

The two-stage PLO findings are the clearest caution from the simulation. In both prevalence domains, heterogeneity alone generated frequent Egger, Begg, and LFK signals, and PLO accounted for most transformation-dependent disagreement. This observation is specific to the continuity-corrected two-stage logit implementation and should not be generalized to one-stage binomial models. More broadly, the prevalence model should be specified before asymmetry results are examined; choosing or switching transformations according to the most favorable asymmetry result would introduce an additional analytical bias.

Where feasible, a count-based random-effects model should be considered for the primary prevalence

synthesis [16]. Among the two-stage scales evaluated in this simulation, PAS produced the most stable asymmetry operating surface. PFT behaved similarly to PAS for asymmetry classification, but concerns regarding inversion and interpretability remain; it is therefore better treated as a sensitivity comparator than as the sole basis for pooled prevalence estimation [13,15]. PLO-based Egger, Begg, and LFK findings require robustness checks using a different analytical scale and explicit review of boundary handling. Study-size-based regressions were more stable under PLO, particularly at 5-50% prevalence, although high heterogeneity still reduced their ability to detect modest effects.

For the practicing anatomist, the relevant choice is not simply between raw and transformed prevalences. Study-level raw prevalence should be reported as the observed number of events divided by the corresponding denominator, but this does not imply that untransformed proportions should be pooled using a normal-normal random-effects model. When event counts and denominators are available, and the model assumptions are appropriate, the preferred primary approach is a one-stage count-based random-effects model, which models the binomial data directly. If a two-stage analysis is required, PAS provided the most stable scale for small-study-effect and asymmetry assessment in the present simulations. PFT should be retained as a sensitivity comparator, whereas continuity-corrected PLO should not be the sole basis for Egger, Begg, or LFK inference and should be checked against another analytical scale. Because untransformed two-stage pooling was not evaluated, these findings do not establish a general ranking of raw versus transformed pooled-prevalence estimators.

Heterogeneity is a modifier, not a nuisance term

Anatomical heterogeneity is often expected rather than anomalous. Population ancestry, imaging or dissection modality, specimen source, classification rules, age, sex, and observer thresholds may all modify true or detected prevalence [7]. I^2 is a relative inconsistency measure that also depends on within-study precision; it cannot distinguish biological diversity from modality, definition, measurement, dependence, or bias. The empirical crosswalk makes this dependence explicit: true logit-scale $\tau = 0.35$ corresponded to median I^2 values of 9.9-46.8% in the rare/low-prevalence platform but approximately 91% in the 5-50% platform. Even when true $\tau = 0$, the upper tail of estimated I^2 reached moderate or substantial values in some cells, whereas sparse rare-event cells could show modest I^2 despite appreciable true logit-scale dispersion. Fixed I^2 categories therefore cannot substitute for the simulation heterogeneity strata. This controlled finding complements an empirical study of 134 prevalence meta-analyses that reported a median I^2 of 96.9% (IQR 90.5-98.7%) and cautioned against arbitrary thresholds [8]. That study established how frequently extreme values occur in prevalence synthesis; the present simulation extends that evidence by demonstrating that the same generated logit-scale heterogeneity can yield markedly different observed distributions across prevalence domains, transformations, and study counts.

The stochastic results add a second qualification. When true random-effect τ was zero, a generated size-prevalence relationship alone produced high I^2, especially in the broader-prevalence platform, because an intercept-only random-effects model treats an unmodeled systematic moderator as residual inconsistency. Authors should therefore report model-scale τ^2 with uncertainty where available, a prediction interval, I^2, study-level prevalence distributions, event yield, and anatomically justified subgroup or meta-regression analyses. Tables *1*-*2* and Figure *13* should be used only for orientation; empirical I^2 should not be converted into, or used to estimate, a simulation true-τ category.

The roadmap downgrades methods under greater heterogeneity for two distinct reasons: some procedures lose calibration, whereas others primarily lose power. These mechanisms should not be conflated. When heterogeneity is high, a negative asymmetry analysis is often inconclusive, while a positive analysis requires stronger triangulation with study characteristics, analytical scale, and sensitivity analyses.

Selective reporting must be assessed separately

The simulation deliberately separated a genuine size-prevalence association from selective dissemination. Under PAS and PFT, selection alone often produced little additional asymmetry despite materially shifting the observed prevalence. This demonstrates a limitation of the target construct: selection can alter the evidence base without creating the effect-size-by-study-size pattern that asymmetry procedures are designed to detect. Conversely, an asymmetry signal may arise from genuine population, modality, design, or measurement differences without selective dissemination.

Risk of missing evidence should therefore be evaluated independently through the review process: protocol and registration searches where relevant, grey and historical literature, conference material, contact with investigators, scrutiny of zero-event findings, comparison of methods with reported outcomes, and identification of overlapping samples. The appropriate conclusion is a graded judgment about the risk of publication or reporting bias, not a declaration that such bias is present or absent on the basis of an asymmetry test alone.

A five-step workflow for anatomists

First, define the observational unit before extraction: person, side, specimen, bone, vessel, branch, foramen,

or another anatomically coherent unit. Record paired or clustered contributions explicitly and use a prevalence model that respects that dependence.

Second, specify the prevalence model before asymmetry analysis. Prefer count-based models when feasible, retain zero-event studies, and document continuity corrections or transformation choices.

Third, characterize heterogeneity anatomically and statistically using model-scale τ^2, its uncertainty when available, a prediction interval, I^2, event yield, study-level prevalence distributions, and prespecified modality, population, or definition analyses. Do not choose a roadmap row from I^2 alone.

Fourth, apply the $K < 10$ gate before formal asymmetry inference. With at least 10 analyzable studies and meaningful size variation, use direct study-size meta-regression for the primary size-related question and triangulate with inverse-n regression and/or Egger-SE. Treat Doi/LFK as descriptive, Begg as supplementary, and trim-and-fill as sensitivity only.

Fifth, interpret any signal within the anatomical evidence structure. Examine whether population, modality, detection threshold, study quality, event yield, dependence, or selective dissemination could explain the pattern; do not convert asymmetry directly into a causal diagnosis of publication bias.

A worked simulated illustration applies these steps without consulting the generating heterogeneity: rare-domain stochastic replicate 1312 contained 20 studies, 462 events, and 20 event-bearing studies. The clean BN fit estimated intercept-only τ at 0.708 and the direct size coefficient at 0.368 (Wald-z $p = 0.055$); the derived t-reference interval was −0.035 to 0.770 (18 degrees of freedom; $p = 0.071$). PAS I^2 was 91.5%, but this does not identify a true-τ roadmap row. Direct-regression p values were 0.103 under PAS and 0.066 under PLO, while PLO Egger signaling was pronounced ($p < 0.001$), unlike PAS ($p = 0.109$). The appropriate reading is an imprecisely estimated size relationship with transformation-dependent asymmetry, not proof of missing studies or absence of a size effect. Heterogeneity uncertainty and anatomical sources such as modality or detection thresholds should guide further assessment; no row is selected mechanically from I^2 or estimated τ. This deliberately selected archived example is illustrative, not external validation; its source key and numerical extract are provided with the reproducibility materials.

Fewer than 10 studies and dependent anatomical data

The decision to pool and the decision to test for small-study effects are separate. A meta-analysis containing fewer than 10 anatomically comparable studies may still provide a useful synthesis, but it offers little basis for formal asymmetry inference. $K = 10$ was the prespecified lower boundary of the present operating analysis and is consistent with widely used guidance; it is not a guarantee of adequate power. Meaningful variation in study size or precision and sufficient event information remain necessary. Below this boundary, the roadmap should not be applied, and the evidence should be described as insufficient for formal asymmetry assessment.

Doi plots and the continuous LFK index may be technically computable with fewer studies, but computability does not establish calibrated operating characteristics. The present simulation did not validate conventional LFK categories below $K = 10$. Funnel or Doi plots may therefore be shown descriptively, but neither their presence nor absence of asymmetry should be used to affirm or exclude publication or reporting bias. The same caution applies to Egger/Begg p values, study-size or inverse-n regression, and trim-and-fill imputations [9].

Asymmetry procedures also cannot diagnose ignored within-person or within-skeleton dependence [24]. Bilateral structures, multiple bones, several branches from one examination, and repeated foramina from one specimen may require paired, multilevel, cluster-robust, beta-binomial, generalized estimating equation, or other correlation-sensitive approaches. Treating left and right structures from the same donor as independent observations inflates precision; that design problem should be addressed during extraction and prevalence modeling rather than sought through a funnel or Doi plot.

Strengths and limitations

The simulation design has three features that aid interpretation: the same evidence bases were analyzed under three transformations; heterogeneity, genuine size-prevalence associations, and selective dissemination were separated by design; and rare/low prevalence was examined alongside the 5-50% range. The controlled anchor and stochastic module provided complementary views of calibration and operating behavior, with the detailed numerical results incorporated in this article.

Several limitations constrain generalization. The study is a simulation stress test rather than an empirical model of every anatomical literature, and the selective-dissemination mechanisms were stylized. Random effects were normal on the logit scale, and the study-size moderator was standardized within each simulated candidate pool. Three two-stage transformations and one one-stage binomial-normal mixed-effects framework were evaluated. Because study event counts were generated from a binomial logistic-normal

structure, the BN-GLMM was correctly specified by design. The direct moderator also used the exact standardized size construct imposed during generation, favoring it for this particular target rather than establishing superiority for nonlinear or differently structured size associations. Its performance therefore represents behavior under model alignment and should not be assumed to generalize to beta-binomial overdispersion, non-normal random-effects distributions, multilevel or clustered dependence, or other forms of model misspecification. Bayesian, cluster-sensitive, and other dependence-aware prevalence models were not evaluated. The LFK calculation used a prespecified manual implementation and the legacy |LFK| > 2 threshold; different implementations or thresholds may change absolute rates. The empirical I^2 crosswalk is specific to the simulated prevalence, study size, event yield, and transformation structures and is not a universal conversion from I^2 to true τ. Formal comparisons were restricted to $K \geqslant 10$, and no external empirical dataset was used to validate the roadmap. The roadmap's 50% detection floor is heuristic, smaller-subgroup Monte Carlo intervals can straddle that floor, and true heterogeneity is unknown in empirical applications. Moreover, t-reference inference did not consistently restore nominal error rates or coverage. The conditional recommendations therefore require sensitivity assessment and external empirical evaluation. Finally, the broader-prevalence extension did not repeat the rare-event reporting-bias uncertainty calculation because that function was designed for low-event omission.

Conclusions

In anatomical prevalence meta-analysis, small-study-effect and asymmetry procedures are influenced jointly by the prevalence model, event yield, transformation, study-size structure, and heterogeneity. Within the evaluated binomial logistic-normal assumptions, one-stage count-based modeling is a primary prevalence-synthesis candidate, and the direct study-size moderator evaluates the size-related pattern imposed in these simulations. This supports a provisional, conditional roadmap rather than general superiority under alternative data-generating structures. Two-stage direct study-size meta-regression provides a transformation-specific comparator; inverse-n regression and Egger-SE provide secondary triangulation; Doi/LFK is best treated as a continuous descriptive index, Begg as supplementary, and trim-and-fill as a sensitivity adjustment. Continuity-corrected two-stage PLO can generate pronounced heterogeneity-related Egger, Begg, and LFK signaling even outside rare-event settings. Monte Carlo uncertainty and the heuristic detection boundary constrain individual ratings. Estimated I^2 cannot be used as a mechanical roadmap selector, and the present simulations do not support formal asymmetry inference when $K < 10$. Most importantly, no asymmetry or moderator procedure diagnoses publication bias. Interpretation should begin with the anatomical data structure and prespecified prevalence model, characterize heterogeneity in clinical and statistical terms, and evaluate the risk of missing evidence separately from size-related asymmetry.

Appendices

The reproducibility materials comprise Clinical_Anatomy_Portable_Reproducibility_FULL.zip and validation_addendum, including readable R source code, dependency manifests, session information, archived validation reports, frozen replicate-level outputs, derived summaries, Monte Carlo uncertainty calculations, and the worked-example extract can be found here: https://drive.google.com/drive/folders/1dggoDo-XjN7MeJ1RhnzSJS3PHZvNAtgh?usp=sharing.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Acquisition, analysis, or interpretation of data: George Triantafyllou, Vasileios Papadopoulos, Maria Piagkou

Critical review of the manuscript for important intellectual content: George Triantafyllou, Maria Piagkou, Aliki Fiska

Concept and design: Vasileios Papadopoulos, Aliki Fiska

Drafting of the manuscript: Vasileios Papadopoulos

Supervision: Maria Piagkou, Aliki Fiska

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.
Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.
Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from

any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

During manuscript preparation, ChatGPT, powered by GPT-5.6 Sol Pro (OpenAI, San Francisco, California, USA), was used to assist with English-language editing and R-code debugging. All AI-assisted suggestions and code modifications were reviewed by the authors. The authors retained full control over the study design, analytical decisions, interpretation of results, and final wording and accept full responsibility for the submitted work.

References

1. Yammine K: Evidence-based anatomy. Clin Anat. 2014, 27:847-52. 10.1002/ca.22397
2. Henry BM, Tomaszewski KA, Walocha JA: Methods of evidence-based anatomy: a guide to conducting systematic reviews and meta-analysis of anatomical studies. Ann Anat. 2016, 205:16-21. 10.1016/j.aanat.2015.12.002
3. Henry BM, Skinningsrud B, Vikse J, et al.: Systematic reviews versus narrative reviews in clinical anatomy: methodological approaches in the era of evidence-based anatomy. Clin Anat. 2018, 31:364-7. 10.1002/ca.23042
4. Tomaszewski KA, Graves MJ, Henry BM, et al.: Surgical anatomy of the sciatic nerve: a meta-analysis . J Orthop Res. 2016, 34:1820-7. 10.1002/jor.23186
5. Solewski B, Lis M, Pękala JR, et al.: The persistent median artery and its vascular patterns: a meta-analysis of 10,394 subjects. Clin Anat. 2021, 34:1173-85. 10.1002/ca.23770
6. Tudose RC, Triantafyllou G, Piagkou M, Rusu MC: Prevalence of the high-riding vertebral artery depends on how it is defined: a meta-analysis with implications for C2 screw fixation. Anat Sci Int. 2026, 10.1007/s12565-026-00954-6
7. Iwanaga J, Walocha J, Ottone NE, Loukas M, Tubbs RS: Extreme heterogeneity reported in published anatomical meta-analyses: a meta-epidemiological study. Clin Anat. 2026, 10.1002/ca.70180
8. Migliavaca CB, Stein C, Colpani V, Barker TH, Ziegelmann PK, Munn Z, Falavigna M: Meta-analysis of prevalence: I(2) statistic and how to deal with heterogeneity. Res Synth Methods. 2022, 13:363-7. 10.1002/jrsm.1547
9. Sterne JA, Sutton AJ, Ioannidis JP, et al.: Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011, 343:d4002. 10.1136/bmj.d4002
10. Hunter JP, Saratzis A, Sutton AJ, Boucher RH, Sayers RD, Bown MJ: In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. J Clin Epidemiol. 2014, 67:897-903. 10.1016/j.jclinepi.2014.03.003
11. Furuya-Kanamori L, Rousou X, Kostoulas P, Doi SA: Examining and interpreting Doi plot asymmetry in meta-analyses of randomized controlled trials. J Evid Based Med. 2025, 18:e70063. 10.1111/jebm.70063
12. Doi SA, Xu C: The Freeman-Tukey double arcsine transformation for the meta-analysis of proportions: recent criticisms were seriously misleading. J Evid Based Med. 2021, 14:259-61. 10.1111/jebm.12445
13. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, Rücker G: Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. Res Synth Methods. 2019, 10:476-83. 10.1002/jrsm.1348
14. Lin L, Xu C: Arcsine-based transformations for meta-analysis of proportions: pros, cons, and alternatives . Health Sci Rep. 2020, 3:e178. 10.1002/hsr2.178
15. Röver C, Friede T: Double arcsine transform not appropriate for meta-analysis . Res Synth Methods. 2022, 13:645-8. 10.1002/jrsm.1591
16. Stijnen T, Hamza TH, Ozdemir P: Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med. 2010, 29:3046-67. 10.1002/sim.4040
17. Egger M, Davey Smith G, Schneider M, Minder C: Bias in meta-analysis detected by a simple, graphical test . BMJ. 1997, 315:629-34. 10.1136/bmj.315.7109.629
18. Begg CB, Mazumdar M: Operating characteristics of a rank correlation test for publication bias . Biometrics. 1994, 50:1088-101. 10.2307/2533446
19. Furuya-Kanamori L, Barendregt JJ, Doi SA: A new improved graphical and quantitative method for detecting bias in meta-analysis. Int J Evid Based Healthc. 2018, 16:195-203. 10.1097/XEB.0000000000000141
20. Duval S, Tweedie R: Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000, 56:455-63. 10.1111/j.0006-341x.2000.00455.x
21. Papadopoulos V, Filippou D, Fiska A: Prevalence of rare anatomic variants - publication bias due to selective reporting in meta-analyses studies. Folia Med (Plovdiv). 2024, 66:795-817. 10.3897/folmed.66.e137881
22. Viechtbauer W: Conducting meta-analyses in R with the metafor package . J Stat Softw. 2010, 36:1-48. 10.18637/jss.v036.i03
23. Schwarzer G, Rücker G, Semaca C: LFK index does not reliably detect small-study effects in meta-analysis: a simulation study. Res Synth Methods. 2024, 15:603-15. 10.1002/jrsm.1714
24. Papadopoulos V, Fiska A: Meta-analysis of paired binary data with unobserved dependence: insights from laterality and bilateralism in anatomy. Stats. 2026, 9:27. 10.3390/stats9020027