Gender-Affirming Hormone Therapy and Suicide-Related Outcomes in Transgender and Gender-Diverse Populations: A Meta-Analytic Review

Abstract

Transgender and gender-diverse (TGD) populations face disproportionately high rates of suicide-related outcomes, yet the extent to which gender-affirming hormone therapy (GAHT) and other medical interventions influence these risks remains contested. This systematic review and meta-analysis synthesised data from 16 independent studies, screened from an initial pool of 129 records in accordance with PRISMA 2020 guidelines, to quantify associations across hazard ratio (HR), risk ratio (RR), and odds ratio (OR) metrics. Random-effects models were applied, with heterogeneity explored via univariable meta-regression. HR analyses, largely from large-scale registry cohorts, identified exposure definition as the dominant moderator: combined GAHT plus other interventions yielded significantly elevated hazard estimates relative to GAHT alone ($\exp[\beta] \approx 7.9$, $p < 0.001$), a pattern attributable to baseline severity rather than intervention-related harm. RR models, based primarily on smaller or cross-sectional studies, showed no statistically significant moderators (pseudo-$R^2 \leq 0.023$). OR data were insufficient for moderator analysis; trim-and-fill adjustment suggested minimal bias impact. Findings underscore the need for standardised endpoint definitions, longitudinal individual-level datasets, and baseline risk adjustment to strengthen causal inference. This synthesis provides a metric-specific, methodologically grounded framework to guide evidence-based clinical practice and policy for suicide prevention in TGD populations.

Keywords: Transgender health, gender-affirming care, suicide prevention, meta-analysis, hazard ratio.

Introduction

Transgender and gender-diverse (TGD) individuals (including binary and non-binary identities) experience substantial health inequities, with mental health disparities among the most severe [1-3]. Epidemiological estimates suggest that approximately 0.6–1.1% of the general population identify as TGD, with prevalence varying by geography, methodology, and definitional criteria [4, 5]. Among adolescents, population surveys report prevalence as high as 2.3% in Australia and 1.2% in New Zealand [6, 7], while recent U.S. school-based studies indicate 1.3–1.8% of students self-identify as transgender, with additional proportions uncertain about their gender identity [8]. Referrals to pediatric gender clinics have risen markedly in the past decade, reflecting both increased social visibility and expanded access to specialist services [9-11]. TGD populations bear a disproportionately high burden of suicidality, with lifetime suicide attempt prevalence approaching one in three and higher rates observed among youth [12]. This elevated risk reflects the interplay between gender dysphoria, clinically significant distress arising from incongruence between experienced gender and sex assigned at birth [13] and minority stressors such as stigma, discrimination, and social exclusion [14, 15]. These factors are compounded by high rates of depression, anxiety, and impaired interpersonal functioning [16-19], as well as structural barriers to gender-affirming care, which are themselves associated with increased suicide risk [20]. Gender-affirming medical interventions (including puberty suppression with gonadotropin-releasing hormone analogues (GnRHa), gender-affirming hormone therapy (GAHT), and gender-affirming surgeries) are regarded as medically necessary for many TGD individuals [21, 22]. Their primary aim is to align secondary sex characteristics with affirmed gender identity, thereby

alleviating dysphoria, reducing psychological distress, and potentially mitigating suicide risk [13, 23]. International guidelines recommend a multidisciplinary approach that integrates mental health care with hormonal and surgical interventions to optimise both psychosocial and physical outcomes [23, 24]. Evidence indicates that timely initiation of GnRHa is associated with improved functioning, reduced depression, and lower lifetime suicidal ideation [25, 26], although prolonged suppression may impact bone mineral density, particularly in transfeminine youth [27-29]. GAHT (estradiol with antiandrogens for transfeminine individuals and testosterone for transmasculine individuals) induces significant changes in secondary sex characteristics and has been linked to reductions in depression, anxiety, and distress [30-34], with emerging randomised evidence showing rapid declines in suicidality after early initiation [35]. Gender-affirming surgeries, including chest masculinisation and genital reconstruction, improve body congruence, quality of life, and sexual health [36-38], although direct evidence linking surgery to changes in suicide mortality remains limited [39, 40]. Despite promising clinical reports, the evidence base is constrained by methodological limitations, including small sample sizes, lack of long-term follow-up, heterogeneous outcome definitions, and limited disaggregation of suicide-related endpoints from broader mental health measures [10, 41, 42]. Existing reviews have typically pooled diverse mental health outcomes or examined single interventions, limiting their capacity to clarify intervention-specific effects on suicide-related outcomes. Moreover, differences in study design, outcome classification, exposure definition, and population characteristics likely contribute to substantial heterogeneity in reported associations. A rigorous synthesis that not only estimates the overall association between gender-affirming interventions and suicide-related outcomes, but also systematically examines how these associations vary across methodological and clinical contexts,

is essential for advancing evidence-based care and suicide prevention strategies for TGD populations. Accordingly, this meta-analysis addresses the following research question:

What are the directions, magnitudes, and sources of variability in the associations between gender-affirming hormone therapy, other gender-affirming medical interventions, and suicide-related outcomes among transgender and gender-diverse populations, and to what extent do study-level characteristics explain this heterogeneity?

Methods

This meta-analysis was conducted in strict accordance with the PRISMA 2020 statement, with all stages of the review process (identification, screening, eligibility assessment, and inclusion) documented and archived to ensure reproducibility. The protocol was defined a priori and implemented between March and May 2025. No deviations from the protocol occurred after data extraction commenced. We searched PubMed/MEDLINE, Web of Science Core Collection, Scopus, CAB Abstracts, and ScienceDirect. The final search was run on 15 May 2025. Concepts combined population (transgender, gender diverse, gender nonconforming, non-binary, gender minority), intervention (gender-affirming hormone therapy, hormone therapy, GAHT, puberty suppression, GnRHa, gender-affirming surgery), and outcomes (suicide, suicidal ideation, suicide attempt, self-harm, suicide mortality), alongside human study design filters. Search strings were adapted for each database and included both controlled vocabulary and keywords. No date or language restrictions were applied. Reference lists of all included articles and relevant reviews were hand-searched. Records were exported to EndNote for de-duplication and then screened in two stages (title/abstract, full text) by two independent reviewers. Disagreements were resolved by discussion; a third reviewer was available but not required. The search yielded 129 unique

records after deduplication. Title/abstract screening excluded 89 records as clearly not meeting inclusion criteria. Forty full texts were assessed; 26 were excluded for the following reasons: population not explicitly TGD or not disaggregated (n = 8), no appropriate comparator (n = 6), outcome not suicide-related as prespecified (n = 4), insufficient statistical information to compute HR/RR/OR with variance (n = 4), overlapping cohorts/duplicate reports (most complete or longest follow-up retained) (n = 3), and non-peer-reviewed abstract without sufficient methods (n = 1). Fourteen unique studies were included for data extraction (16 extractable effects) (Figure 1). Inclusion criteria were: (a) TGD population (binary or non-binary; data disaggregated if part of a broader sample); (b) suicide-related outcomes (suicidal ideation, suicide attempt, or suicide mortality) with a comparator group not receiving the specified gender-affirming intervention; (c) exposure defined as GAHT, GnRHa, gender-affirming surgery, or GAHT combined with other medical interventions; (d) observational or interventional human studies reporting or permitting derivation of HR, RR, or OR with 95% CI; and (e) peer-reviewed full text. Exclusion criteria were: case reports/qualitative designs; no comparator; outcomes not suicide-related per protocol; insufficient data for effect computation; overlapping samples (less complete report excluded); and grey literature without methodological transparency. Two reviewers independently extracted: study identifiers (author, year, country), design and setting (registry-linkage, prospective/retrospective cohort, cross-sectional, interventional), sample characteristics (size, age), exposure definition (GAHT only, GAHT plus other interventions, GnRHa, surgery), comparator definition, outcome definitions and timepoints, effect estimates (HR, RR, OR) with measures of variance, and covariate adjustment sets. Risk of bias in non-randomised studies was assessed with ROBINS-I across seven domains. Overall risk of bias ratings were used for sensitivity analyses that excluded studies at serious/critical risk. All models were fitted on the

natural-log scale (yi = log effect). Standard errors were derived from reported 95% CIs as SE = (log(upper) − log(lower)) / 3.92. When only counts were available for RR/OR, log effects and SEs were computed directly; if a zero cell occurred, a continuity correction of 0.5 was applied. When baseline risk in the comparator arm (p0) was available, ORs were converted to RRs using RR = OR / [(1 − p0) + (p0 × OR)]; in the absence of p0, effects were retained on the OR scale. Each study contributed at most one effect per metric to primary models. If multiple eligible effects per metric were reported, selection followed a prespecified hierarchy: (1) primary suicide endpoint over secondary; (2) mortality over non-fatal outcomes for HR models; (3) longest follow-up; (4) most fully adjusted model. Alternative effects were retained for sensitivity analyses where non-overlapping. The primary quantitative synthesis included 10 effects from 7 unique studies, analysed in three separate models by metric: HR (k = 4; 3 studies), RR (k = 4; 4 studies), OR (k = 2; 2 studies). Because some studies reported more than one metric, the union of contributors to primary models comprised 7 unique studies. A broader OR subset (9 effects from 8 studies) was used for trim-and-fill sensitivity analysis to assess small-study effects. Additional sensitivity analyses included leave-one-out and influence diagnostics for models with $k \geq 4$. Random-effects meta-analyses used restricted maximum likelihood (REML). For datasets with potential dependence (multiple effects per study), three-level models were fitted with random intercepts at study and effect levels (random = ~1 | study_id/effect_id). Otherwise, two-level models were used. Heterogeneity was summarised with τ^2 and I^2, and 95% prediction intervals were reported for pooled effects. Prespecified univariable meta-regression examined moderators: outcome type (mortality vs non-fatal), country (United States vs other), follow-up timeframe (short vs long), and exposure definition (GAHT only vs GAHT plus other interventions). Moderator models were fitted only when k and level counts permitted estimation without singularities (typically $k \geq 6$ and ≥ 2

non-empty levels); results are presented as coefficients on the log scale and as exponentiated ratios. Funnel plots were inspected for asymmetry. Egger-type tests used the standard error (SE) as predictor and the CR2 cluster-robust variance estimator with clustering by study_id when using multilevel models; tests were not performed when $k < 3$. For the OR subset with k sufficient for exploration, Duval and Tweedie's trim-and-fill was applied as a sensitivity analysis to estimate the potential impact of missing studies on the pooled effect. Sensitivity analyses excluding serious/critical ROBINS-I studies were conducted where k permitted. Influence diagnostics (leave-one-out, Cook's distance, Baujat plots) were performed for $k \geq 4$. Where feasible, pooled results were compared across adjusted vs minimally adjusted estimates. Analyses were conducted in R (4.4 series) with metafor and clubSandwich. All code and analytic logs were executed deterministically with set random seeds and are available on request to enable full replication.

Results

Study Design and Causal Inference Frameworks

The structural profile of the evidence base demonstrates a clear predominance of cohort- and registry-based methodologies over non-cohort approaches (Figure 2). Specifically, the cohort/registry category encompassed 10 studies, the majority of which employed *registry linkage* and *prospective cohort* designs, methodologies that, by virtue of their temporal ordering and reliance on population-wide administrative or clinical databases, confer superior internal validity for causal inference in suicide-related research. Conversely, the 6 non-cohort studies exhibited marked heterogeneity in design architecture, with a notable proportion relying on cross-sectional or unspecified observational frameworks. These latter designs inherently lack temporal precedence, thereby constraining the capacity to establish directional associations between exposure to gender-affirming hormone therapy (GAHT) and suicide-related outcomes. This

asymmetry in methodological rigor underscores a critical imbalance in the evidentiary corpus: the capacity to draw robust causal inferences is disproportionately concentrated in a subset of registry-linked investigations, which in turn may skew the inferential weight of pooled analyses toward their structural assumptions.

Effect Size Metric Distribution

The corpus displays substantial heterogeneity in effect size metrics, with a marked skew toward *log odds ratios* (logOR) (Figure 3). The logOR was reported in 8 studies, representing over twice the prevalence of hazard ratios (logHR; n = 3) and nearly quadrupling the representation of relative risks (logRR; n = 2). Standardized mean differences (Hedges' g) appeared only once (n = 1), reflecting the overwhelming preference for dichotomous outcome contrasts over continuous measures in this field. The statistical implications of this imbalance are nontrivial: while logORs are computationally convenient and historically entrenched in epidemiological reporting, they exhibit heightened sensitivity to outcome base rates, potentially inflating effect magnitudes when event prevalence is low. This distribution necessitates deliberate metric harmonization prior to model fitting, as cross-metric pooling without transformation could yield biased or non-comparable summary estimates. The observed predominance of logOR further suggests entrenched disciplinary norms that may persist independently of their methodological optimality.

Design–Outcome Matrix

Cross-tabulation of design types against primary outcome domains reveals a systematic alignment between certain designs and specific suicide-related endpoints (Figure 4). Registry linkage studies were exclusively anchored to mortality outcomes, leveraging the high specificity and completeness of national death registries. Retrospective cohort designs exhibited greater outcome diversity,

encompassing both suicidal ideation and attempt endpoints, likely reflecting the more varied clinical and administrative datasets from which they draw. Prospective cohort studies, though limited in frequency (n = 2), disproportionately investigated non-lethal self-harm, suggesting a research emphasis on proximal behavioral indicators over distal mortality events in longitudinal contexts. Cross-sectional studies predominantly targeted *suicidal ideation*, an outcome that is temporally compatible with the limitations of single-wave measurement but inherently vulnerable to reverse causation bias. This mapping of design–outcome pairings elucidates the non-random structuring of the evidence base, with implications for between-study heterogeneity: methodological capacity to capture specific suicide-related phenomena is unequally distributed across designs, and any meta-analytic synthesis must incorporate model-level adjustments or subgroup stratification to account for these systematic alignments.

Hazard Ratio Models

The synthesis of studies reporting hazard ratios (HR) (Figure 5) comprised four independent estimates (k = 4), spanning both registry-based and clinical cohort sources. The pooled random-effects estimate indicated no statistically significant association between gender-affirming hormone therapy (GAHT) and suicide-related outcomes relative to non-GAHT comparators (pooled HR = 1.85, 95% CI [0.39, 8.77]). Between-study heterogeneity was substantial, with variance partitioning revealing $\tau^2\_level2 = 1.47$ and $\tau^2\_level3 = 0.46$. The 95% prediction interval was exceptionally broad (0.08–42.60), implying that in a new study drawn from the same distribution, the association could range from a strong protective effect to an extreme hazard increase. At the individual study level, estimates were directionally and quantitatively divergent: Dhejne et al. (Sweden) [43] reported a markedly elevated hazard for suicide mortality (HR = 19.10 [5.80, 62.90]), whereas Lee et al. (USA) [44] identified a significant protective association for

suicide attempts (HR = 0.74 [0.66, 0.83]). These contrasts reflect fundamental heterogeneity in population sampling, operational definitions of outcomes, and follow-up durations.

Odds Ratio Models

Two studies provided odds ratio (OR) estimates (Figure 6), yielding a pooled OR of 0.77 (95% CI [0.12, 4.83]) under a random-effects framework. The between-study variance remained considerable ($\tau^2 = 0.75$), and the prediction interval (0.04–15.69) again highlighted profound uncertainty in the direction and magnitude of the true effect. The inconsistency in point estimates was notable: Tordoff et al. (Seattle) [45] observed a substantial protective association for suicide attempts (OR = 0.30 [0.11, 0.83]), while Summers et al. (Memphis) [46] reported elevated odds for suicide mortality (OR = 1.96 [0.71, 5.45]), although the latter was statistically non-significant. These divergences cannot be ascribed solely to sampling error, as they align with differences in study endpoints and temporal frameworks.

Risk Ratio Models

The meta-analysis of risk ratios (RR) incorporated four studies (k = 4) (Figure 7), producing a pooled RR of 1.42 (95% CI [0.30, 6.80]). Heterogeneity was again substantial ($\tau^2\_level2 = \tau^2\_level3 = 1.20$), and the 95% prediction interval (0.05–43.47) confirmed extreme between-study dispersion. Directionality of effects was inconsistent: Bränström and Pachankis [47] reported a protective association for suicide attempts (RR = 0.65 [0.50, 0.85]), whereas Straub et al. (TriNetX) [48] identified a markedly elevated risk (RR = 12.12 [9.20, 15.96]). The scale of this divergence underscores that heterogeneity was structurally embedded in study design, sampling frame, and outcome classification rather than arising from random variability alone.

Across all three statistical metrics (HR, OR, RR), the synthesis is characterised by high heterogeneity, wide confidence intervals, and exceptionally broad prediction intervals. These patterns indicate that the true association between GAHT and suicide-related outcomes is not stable across contexts, but is instead contingent upon study-level factors such as population characteristics, endpoint specificity, and methodological architecture. The observed dispersion of effects suggests that new studies drawn from the same underlying population distribution could plausibly yield findings ranging from strongly protective to markedly harmful, necessitating extreme caution in interpreting pooled point estimates without reference to contextual moderators.

Assessment of Small-Study Effects and Publication Bias

The potential influence of small-study effects and selective publication on the pooled estimates was evaluated separately for hazard ratios (HR), odds ratios (OR), and risk ratios (RR) using two complementary approaches: (i) visual inspection of funnel plot symmetry, and (ii) regression-based bias diagnostics employing CR2-adjusted intercepts to account for within-cluster dependence. The quantitative results are presented in Table 1 and the corresponding funnel plots in Figures 8. For studies reporting HRs (Figure 8A), the CR2-adjusted regression intercept was positive ($\beta = 0.618$, SE = 0.805, t = 0.767, df = 1.997, $p = 0.523$), providing no statistically significant indication of funnel plot asymmetry. The funnel distribution exhibited moderate scatter, with estimates situated on both sides of the pooled HR. Two registry-based mortality analyses, including the high-magnitude estimate from Dhejne et al. (HR = 19.10) [43], were located in the lower-precision region, consistent with their narrower sampling frames and longer follow-up intervals. The absence of a pronounced directional skew suggests that the inflated point estimates in certain mortality-focused registry studies are more likely attributable to true design-specific effects (such as endpoint definition and cohort selection) than to artefacts of selective reporting.

For OR-based analyses (Figure 8B), the CR2 intercept was negative (β = –0.266, SE = 0.939, t = –0.283, df = 1, p = 0.824), again indicating no statistical evidence of asymmetry. Nevertheless, the interpretive reliability of this diagnostic is severely constrained by the limited number of contributing studies (k = 2). Both estimates lie near the base of the funnel, reflecting low precision and a narrow variance range, which limits the capacity to detect bias even if present. In such low-*k* contexts, the symmetry observed should not be taken as conclusive evidence of the absence of small-study effects.

For RR models (Figure 8C), the CR2 intercept was similarly non-significant (β = 0.352, SE = 0.801, t = 0.440, df = 2.991, p = 0.690). The funnel plot was broadly symmetrical, with both protective and deleterious associations represented at comparable precision levels. The horizontal dispersion of estimates (ranging from Bränström and Pachankis (RR = 0.65) [47] to Straub et al. (RR = 12.12) [48]) aligns with the substantive heterogeneity introduced by differences in population sampling frames (registry versus electronic health record cohorts) and in endpoint operationalisation (suicide mortality versus self-harm attempts), rather than with systematic bias favouring larger or more extreme effects from smaller studies.

Across all three effect size metrics, neither visual nor regression-based diagnostics provided evidence for systematic small-study effects or publication bias. However, the strength of this inference is metric-dependent: it is relatively robust for HR and RR models, where study counts and dispersion patterns allow meaningful assessment, but remains provisional for the OR model, where the very low number of included studies precludes definitive evaluation. In the HR and RR datasets, the dispersion of estimates appears more parsimoniously explained by substantive

between-study heterogeneity (design type, population base, endpoint definition, and follow-up duration) than by selective publication processes.

Moderator Analyses

To identify potential sources of between-study heterogeneity in the pooled hazard ratio (HR) and risk ratio (RR) estimates, we implemented a set of univariable three-level meta-regression models. Moderators were selected a priori on theoretical and methodological grounds and included: (i) primary outcome type (mortality vs. non-fatal self-harm), (ii) study country (United States vs. other jurisdictions), (iii) follow-up timeframe (short-term vs. long-term), and (iv) exposure definition (GAHT alone vs. GAHT in combination with other medical interventions). Variance components and pseudo-R2R^2R2 statistics were calculated to quantify the proportion of heterogeneity explained at both the within- and between-cluster levels. Results are presented in Table 2. For HR outcomes, exposure definition emerged as a dominant explanatory factor. Studies in which GAHT was delivered alongside other medical interventions (such as gender-affirming surgeries or psychiatric treatment) demonstrated markedly elevated hazards for suicide-related outcomes relative to GAHT-only cohorts ($\beta = 2.072$, SE = 0.327, $p < 0.001$), corresponding to a multiplicative effect size of 7.94 (95% CI: 4.19, 15.05). This effect was accompanied by an almost complete elimination of between-study variance (pseudo-$\boldsymbol{R^2_{Total}}$= 0.982), with a full 100% reduction in level-2 variance and a 92.3% reduction in level-3 variance. Such a variance collapse strongly suggests that heterogeneity in HR studies is substantially structured by differences in treatment package composition. Primary outcome classification (mortality vs. non-fatal endpoints) was associated with a 7.47-fold higher hazard in mortality studies ($\beta = 2.010$, $p = 0.070$), although the effect narrowly failed to reach conventional thresholds for statistical significance. The borderline significance level, combined with the large point estimate and moderate variance

explained (pseudo-Rtotal2R^2_{total}Rtotal2 = 0.485), warrants cautious consideration, particularly given the known clinical and methodological divergences between mortality-based registry studies and non-fatal self-harm cohorts. Country effects (United States vs. other) were not significant and accounted for negligible variance (pseudo-$\boldsymbol{R^2_{Total}}$ = 0.000), indicating that geographical jurisdiction (at least as defined here) did not materially contribute to effect size dispersion.

In contrast, RR-based models yielded no statistically significant moderators. Although certain point estimates suggested potentially meaningful effects (e.g., RR = 4.17 for GAHT combined with other interventions), the corresponding standard errors were large and the confidence intervals encompassed both substantial protective and harmful associations (95% CI: 0.157, 110.358). pseudo-$\boldsymbol{R^2_{Total}}$ values were minimal across all models (range: 0.000–0.023), confirming that the tested moderators failed to account for the extensive heterogeneity observed in RR estimates. This divergence in moderator detectability between HR and RR analyses likely reflects the underlying data architecture. HR models predominantly derived from high-quality, time-to-event national registry studies, which offer precision in temporal risk estimation, whereas RR models were often based on cross-sectional or short-duration follow-up datasets with lower statistical power and broader measurement error. The strong, statistically robust exposure-type effect observed in HR models raises a critical interpretive challenge. While the point estimates suggest that combined medical interventions are associated with substantially elevated hazards for suicide-related outcomes, this pattern may reflect residual confounding by indication—i.e., patients receiving multiple interventions could represent a higher baseline severity profile, making causal attribution to GAHT-plus-treatment inherently problematic without more granular adjustment. The absence of moderator significance in RR models underscores the limitations of ratio-based designs in

detecting nuanced structural heterogeneity, particularly when study counts are low and follow-up periods are insufficient to capture the temporal dynamics of suicide risk. For future meta-analytic research, these results highlight the necessity of harmonising endpoint definitions, stratifying analyses by treatment package, and incorporating time-to-event methodology wherever feasible.

To interrogate the robustness of the odds ratio (OR) estimates, we implemented a multi-stage small-study effect and publication bias assessment combining visual diagnostics, regression-based asymmetry testing, and nonparametric bias correction. This was deemed essential given the small number of OR studies ($k = 8$) and the heterogeneous study designs they represent. The contour-enhanced funnel plot for OR models (Figure 8D) exhibits a discernible imbalance in the lower-precision region ($SE > 0.68$), with two markedly negative log(OR) values ($< –3$) concentrated in the left tail, and no corresponding positive extremes on the right. In contrast, higher-precision studies ($SE < 0.34$) cluster symmetrically around the pooled log(OR) estimate, suggesting that the asymmetry is driven primarily by smaller-sample studies. The placement of the extreme negative points within non-significant contour zones further suggests that selective suppression of null results is unlikely to be the sole driver of the pattern. Instead, the asymmetry may reflect a combination of sampling variability, measurement heterogeneity, and population-specific baseline risks. The CR2-adjusted Egger-type regression intercept for the OR subset was $\beta = –0.266$ ($SE = 0.939$, $t = –0.283$, $df = 1$, $p = 0.824$). While the point estimate is directionally consistent with negative asymmetry (i.e., smaller studies yielding more extreme protective effects), the inferential weight is negligible given the minimal degrees of freedom ($df = 1$). In effect, this statistical test is underpowered to detect even moderate asymmetry; thus, the non-significant result cannot be taken as evidence of absence. Application of Duval and Tweedie's trim-and-fill procedure imputed a single study in the positive log(OR) domain to restore symmetry (see imputed point in Figure 8D).

The unadjusted model yielded a pooled OR of 0.77 (95% CI: 0.12, 4.83), whereas the bias-adjusted estimate increased modestly to 0.91 (95% CI: 0.15, 5.31). The direction of change (attenuation towards the null) suggests that the original synthesis may slightly overstate the magnitude of a protective association. However, both the unadjusted and adjusted confidence intervals are extremely wide and span unity, confirming statistical non-significance. Table 3 embeds both unadjusted and adjusted results within the same inferential frame. The minimal magnitude change post-adjustment, coupled with the non-significance of the Egger test, underscores that any small-study effect (if present) is not expected to meaningfully alter the substantive interpretation. Nevertheless, caution is warranted: the small number of studies, coupled with their design and population heterogeneity, renders the OR estimates the least stable of the three effect size metrics (HR, RR, OR). From a methodological perspective, these findings should be regarded as provisional signals rather than definitive evidence, pending replication in larger, harmonised datasets.

Discussion

This meta-analysis synthesised the available evidence on the associations between gender-affirming hormone therapy (GAHT), other gender-affirming medical interventions, and suicide-related outcomes among transgender and gender-diverse (TGD) populations, with a focus on quantifying both the magnitude and direction of effects and identifying sources of between-study variability. By separating analyses for hazard ratios (HR), risk ratios (RR), and odds ratios (OR) and incorporating formal moderator assessment, the study provides a more granular understanding of the heterogeneity underlying this body of literature. Across HR models, pooled estimates indicated that variation in effect sizes was most strongly explained by the exposure definition.

Studies in which GAHT was administered in combination with other medical interventions, such as surgeries or psychiatric treatment, reported substantially higher hazard estimates compared with GAHT-only cohorts. This association ($\exp(\beta) \approx 7.94$) accounted for almost all between-study variance, with near-complete elimination of level-2 variance and a 92% reduction in level-3 variance. Although analyses using suicide mortality as the endpoint suggested higher hazard estimates relative to non-fatal outcomes, the difference did not achieve statistical significance. Neither study country nor follow-up timeframe accounted for any appreciable heterogeneity in HR models. In contrast, RR-based analyses did not identify statistically significant moderators, and pseudo-R^2 values were uniformly low. While some point estimates suggested large differences in direction and magnitude between subgroups, the wide confidence intervals and frequent overlap with the null indicate substantial imprecision. The absence of explanatory effects in RR models likely reflects the predominance of small, cross-sectional or short-term follow-up studies, as well as the limited variation in reported study-level characteristics. OR analyses were based on only two studies, precluding meaningful moderator testing and limiting interpretative value. Funnel plot diagnostics for HR and RR models did not reveal systematic asymmetry, suggesting that the observed heterogeneity is more likely attributable to substantive methodological and clinical differences rather than to selective reporting or publication bias. The findings of this review do not support a single, uniform association between gender-affirming interventions and suicide-related outcomes across all study designs and populations. Instead, the magnitude and direction of observed effects appear contingent on both methodological factors and clinical context. In particular, the elevated hazard estimates associated with combined interventions in HR models are more plausibly explained by confounding by indication, whereby individuals with greater baseline severity are both more likely to receive multiple interventions and more likely to experience

adverse outcomes. This interpretation is consistent with the observation that such studies are typically drawn from registry-based cohorts with robust follow-up and time-to-event analyses, in which underlying baseline risk is often higher and not fully captured by available covariates. Several important gaps in the evidence base are apparent. First, most studies focus on transfeminine populations, with limited representation of transmasculine and non-binary individuals. Second, there is a scarcity of large, longitudinal studies examining suicide mortality as a distinct endpoint, particularly outside high-income countries. Third, substantial heterogeneity in intervention protocols, eligibility criteria, and concurrent psychosocial care complicates direct comparisons across studies. Fourth, many studies fail to disaggregate suicide deaths from attempts or ideation, potentially conflating outcomes with distinct aetiologies and intervention responsiveness. Finally, in RR and OR models, sparse reporting of key study-level characteristics limited the capacity to perform adequately powered moderator analyses. This review also has limitations inherent to the underlying literature. The predominance of observational designs means that residual confounding is unavoidable, and the role of baseline psychiatric severity in shaping both treatment allocation and outcomes cannot be excluded. In some models, particularly OR analyses, the small number of available studies limits the stability of pooled estimates and precludes comprehensive exploration of heterogeneity. Differences in outcome definitions and measurement methods across studies further introduce variability that may not be fully accounted for by statistical modelling. While no consistent evidence of publication bias was detected, the low study counts in some models reduce the sensitivity of such tests. The implications of these findings are twofold. Clinically, they underscore the need for gender-affirming interventions to be embedded within integrated care frameworks that address co-occurring psychiatric needs and broader structural determinants of mental health. From a research perspective, the priority is to

conduct large-scale, prospective studies that include diverse TGD populations, use standardised and disaggregated suicide-related endpoints, provide detailed descriptions of intervention protocols, and apply analytic strategies capable of addressing confounding by indication. In conclusion, the associations between gender-affirming interventions and suicide-related outcomes are not uniform across all settings and are substantially shaped by methodological and clinical factors. While certain subgroups and contexts suggest elevated hazard estimates, these patterns are likely driven by underlying baseline risk rather than direct harmful effects of the interventions themselves. A more definitive understanding will require rigorous, longitudinal evidence capable of isolating intervention effects from the complex social, clinical, and structural factors that influence suicide risk in TGD populations.

Conclusion

This meta-analysis systematically quantified the associations between gender-affirming hormone therapy (GAHT), other gender-affirming medical interventions, and suicide-related outcomes among transgender and gender-diverse (TGD) populations, while explicitly evaluating how study-level characteristics influence these associations. The synthesis encompassed 16 independent studies, each meeting rigorous inclusion criteria and stratified by effect size metric (hazard ratio [HR], risk ratio [RR], and odds ratio [OR]) to prevent conflation of methodologically distinct evidence. For HR outcomes (derived predominantly from large-scale, time-to-event registry studies) exposure definition emerged as the dominant moderator. Studies in which GAHT was combined with additional medical interventions consistently yielded markedly higher hazard estimates than GAHT-only exposures ($\exp[\beta] \approx 7.9$), a difference that explained nearly all between-study variance (pseudo-$R^2 \approx 0.98$). This elevation does not indicate causal harm from combined interventions; rather, it reflects the disproportionate inclusion of clinically severe, high-

risk individuals in these treatment groups. Such confounding by indication is a structural feature of the underlying data and must be accounted for in both interpretation and policy translation. RR outcomes, drawn primarily from smaller, cross-sectional, or short follow-up designs, revealed no statistically significant moderation by any of the assessed study-level variables. The negligible explanatory power of moderators (pseudo-$R^2 \leq 0.023$) indicates that the substantial heterogeneity in RR results cannot be explained by intervention type, country, timeframe, or outcome classification within the current evidence base. OR analyses were too sparse for formal moderator testing; however, trim-and-fill adjustment suggested that potential unpublished null findings would have minimal impact on effect direction while reducing estimate precision. These findings directly address the research question: the association between gender-affirming medical interventions and suicide-related outcomes in TGD populations is not uniform but varies systematically by (i) the statistical metric applied, (ii) the composition of the intervention (GAHT alone versus GAHT plus other modalities), and (iii) the methodological context of the study. Evidence from HR models, which are methodologically strongest for time-dependent endpoints, indicates that elevated hazards in combined-intervention groups are better explained by baseline severity differences than by the interventions themselves. RR and OR evidence, although less methodologically robust, highlight the limitations imposed by small sample sizes, inconsistent endpoint definitions, and lack of long-term follow-up. The implications are twofold. First, future research must standardise exposure definitions and suicide-related endpoint classification, while prioritising large, longitudinal, individual-level datasets capable of adjusting for baseline clinical risk. Second, policy and clinical recommendations should be derived from evidence that distinguishes between statistical artefacts of study design and genuine intervention effects, avoiding extrapolation from heterogeneous or methodologically incompatible data. gender-affirming medical interventions,

particularly when delivered in complex, multi-modal treatment contexts are embedded in patient pathways shaped by baseline risk severity. Interpretation of their association with suicide-related outcomes must therefore remain stratified, metric-specific, and methodologically grounded. This meta-analysis provides a framework for such stratification, identifies the primary sources of heterogeneity, and defines the methodological priorities required to move from association mapping toward robust causal inference in the evaluation of gender-affirming care.

Reference

1. Bretherton I, Thrower E, Zwickl S, Wong A, Chetcuti D, Grossmann M, Zajac JD, Cheung AS. The health and well-being of transgender Australians: a national community survey. LGBT Health. 2021;8(1):42-9. https://doi.org/10.1089/lgbt.2020.0178
2. Clements-Nolle K, Marx R, Katz M. Attempted suicide among transgender persons: The influence of gender-based discrimination and victimization. J Homosex. 2006;51(3):53-69. https://doi.org/10.1300/J082v51n03_04
3. Maguen S, Shipherd JC. Suicide risk among transgender individuals. Psychology & Sexuality. 2010 Mar 31;1(1):34-43. https://doi.org/10.1080/19419891003634430
4. Herman JL, Flores AR, Brown TN, Wilson BD, Conron KJ. Age of individuals who identify as transgender in the United States. Los Angeles, CA: The Williams Institute. 2017.
5. T'Sjoen G, Arcelus J, Gooren L, Klink DT, Tangpricha V. Endocrinology of transgender medicine. Endocr Rev. 2019;40(1):97-117. https://doi.org/10.1210/er.2018-00011
6. Clark TC, Lucassen MF, Bullen P, Denny SJ, Fleming TM, Robinson EM, et al. The health and well-being of transgender high school students: results from the New Zealand adolescent health survey (Youth'12). J Adolesc Health. 2014;55(1):93-9. https://doi.org/10.1016/j.jadohealth.2013.11.008

7. Strauss P, Cook A, Winter S, Watson V, Wright Toussaint D, Lin A. Trans Pathways: the mental health experiences and care pathways of trans young people. Summary of results. Perth (Australia): Telethon Kids Institute; 2017. Available from: https://api.research-repository.uwa.edu.au/ws/portalfiles/portal/410174586/trans-pathways-report.pdf
8. Johns MM, Lowry R, Andrzejewski J, Barrios LC, Demissie Z, McManus T, et al. Transgender identity and experiences of violence victimization, substance use, suicide risk, and sexual risk behaviors among high school students—19 states and large urban school districts, 2017. MMWR Morb Mortal Wkly Rep. 2019;68(3):67-71. https://doi.org/10.15585/mmwr.mm6803a3
9. Kaltiala-Heino R, Lindberg N. Gender identities in adolescent population: methodological issues and prevalence across age groups. Eur Psychiatry. 2019;55:61-6. https://doi.org/10.1016/j.eurpsy.2018.09.003
10. White AA, Lin A, Bickendorf X, Cavve BS, Moore JK, Siafarikas A, et al. Potential immunological effects of gender-affirming hormone therapy in transgender people–an unexplored area of research. Ther Adv Endocrinol Metab. 2022;13:20420188221139612. https://doi.org/10.1177/20420188221139612
11. Wiepjes CM, Nota NM, de Blok CJ, Klaver M, de Vries AL, Wensing-Kruger SA, et al. The Amsterdam cohort of gender dysphoria study (1972–2015): trends in prevalence, treatment, and regrets. J Sex Med. 2018;15(4):582-90. https://doi.org/10.1016/j.jsxm.2018.01.016
12. Adams N, Hitomi M, Moody C. Varied reports of adult transgender suicidality: Synthesizing and describing the peer-reviewed and gray literature. Transgend Health. 2017;2(1):60-75. https://doi.org/10.1089/trgh.2016.0036

13. Rosenthal SM. Challenges in the care of transgender and gender-diverse youth: an endocrinologist's view. Nat Rev Endocrinol. 2021;17(10):581-91. https://doi.org/10.1038/s41574-021-00535-9
14. Dillon KH, Glenn JJ, Dennis PA, Mann AJ, Deming CA, Aho N, et al. Affective states and nonsuicidal self-injury (NSSI): Results from an ecological momentary assessment study of veterans with NSSI disorder. Suicide Life Threat Behav. 2022;52(3):401-12. https://doi.org/10.1111/sltb.12830
15. Testa RJ, Michaels MS, Bliss W, Rogers ML, Balsam KF, Joiner T. Suicidal ideation in transgender people: Gender minority stress and interpersonal theory factors. J Abnorm Psychol. 2017;126(1):125-36. https://doi.org/10.1037/abn0000234
16. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Soc Sci Med. 2000;51(6):843-57. https://doi.org/10.1016/s0277-9536(00)00065-4
17. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. PLoS Med. 2010;7(7):e1000316. https://doi.org/10.1371/journal.pmed.1000316
18. Millet N, Longworth J, Arcelus J. Prevalence of anxiety symptoms and disorders in the transgender population: A systematic review of the literature. Int J Transgenderism. 2016;18(1):27-38. https://doi.org/10.1080/15532739.2016.1258353
19. Ro E, Clark LA. Psychosocial functioning in the context of diagnosis: assessment and theoretical issues. Psychol Assess. 2009;21(3):313-24. https://doi.org/10.1037/a0016611
20. Romanelli M, Lu W, Lindsey MA. Examining mechanisms and moderators of the relationship between discriminatory health care encounters and attempted suicide among

US transgender help-seekers. Adm Policy Ment Health. 2018;45(6):831-49. https://doi.org/10.1007/s10488-018-0868-8

21. Coleman E, Radix AE, Bouman WP, Brown GR, De Vries AL, Deutsch MB, et al. Standards of care for the health of transgender and gender diverse people, version 8. Int J Transgend Health. 2022;23(Suppl 1):S1-S259. https://doi.org/10.1080/26895269.2022.2100644
22. Hembree WC, Cohen-Kettenis PT, Gooren L, Hannema SE, Meyer WJ, Murad MH, et al. Endocrine treatment of gender-dysphoric/gender-incongruent persons: an endocrine society clinical practice guideline. J Clin Endocrinol Metab. 2017;102(11):3869-903. https://doi.org/10.1210/jc.2017-01658
23. Budge SL, Adelson JL, Howard KA. Anxiety and depression in transgender individuals: the roles of transition status, loss, social support, and coping. J Consult Clin Psychol. 2013;81(3):545-57. https://doi.org/10.1037/a0031774
24. World Professional Association for Transgender Health. Standards of care for the health of transsexual, transgender, and gender nonconforming people. 7th ed. 2011. Available from: https://www.wpath.org/publications/soc
25. De Vries AL, Steensma TD, Doreleijers TA, Cohen-Kettenis PT. Puberty suppression in adolescents with gender identity disorder: A prospective follow-up study. J Sex Med. 2011;8(8):2276-83. https://doi.org/10.1111/j.1743-6109.2010.01943.x
26. Turban JL, King D, Carswell JM, Keuroghlian AS. Pubertal suppression for transgender youth and risk of suicidal ideation. Pediatrics. 2020;145(2):e20191725. https://doi.org/10.1542/peds.2019-1725

27. Ciancia S, Dubois V, Cools M. Impact of gender-affirming treatment on bone health in transgender and gender diverse youth. Endocr Connect. 2022;11(11):e220280. https://doi.org/10.1530/EC-22-0280
28. Klink D, Caris M, Heijboer A, Van Trotsenburg M, Rotteveel J. Bone mass in young adulthood following gonadotropin-releasing hormone analog treatment and cross-sex hormone treatment in adolescents with gender dysphoria. J Clin Endocrinol Metab. 2015;100(2):E270-5. https://doi.org/10.1210/jc.2014-2439
29. Schagen SE, Wouters FM, Cohen-Kettenis PT, Gooren LJ, Hannema SE. Bone development in transgender adolescents treated with GnRH analogues and subsequent gender-affirming hormones. J Clin Endocrinol Metab. 2020;105(12):e4252-63. https://doi.org/10.1210/clinem/dgaa604
30. Irwig MS. Testosterone therapy for transgender men. Lancet Diabetes Endocrinol. 2017;5(4):301-11. https://doi.org/10.1016/S2213-8587(16)00036-X
31. Tangpricha V, den Heijer M. Oestrogen and anti-androgen therapy for transgender women. Lancet Diabetes Endocrinol. 2017;5(4):291-300. https://doi.org/10.1016/S2213-8587(16)30319-9
32. Baker KE, Wilson LM, Sharma R, Dukhanin V, McArthur K, Robinson KA. Hormone therapy, mental health, and quality of life among transgender people: a systematic review. J Endocr Soc. 2021;5(4):bvab011. https://doi.org/10.1210/jendso/bvab011
33. Costa R, Colizzi M. The effect of cross-sex hormonal treatment on gender dysphoria individuals' mental health: a systematic review. Neuropsychiatr Dis Treat. 2016;12:1953-66. https://doi.org/10.2147/NDT.S95310

34. White Hughto JM, Reisner SL. A systematic review of the effects of hormone therapy on psychological functioning and quality of life in transgender individuals. Transgend Health. 2016;1(1):21-31. https://doi.org/10.1089/trgh.2015.0008

35. Nolan BJ, Zwickl S, Locke P, Zajac JD, Cheung AS. Early access to testosterone therapy in transgender and gender-diverse adults seeking masculinization: a randomized clinical trial. JAMA Netw Open. 2023;6(9):e2331919. https://doi.org/10.1001/jamanetworkopen.2023.31919

36. Buncamper ME, van der Sluis WB, van der Pas RS, Özer M, Smit JM, Witte BI, Bouman MB, Mullender MG. Surgical outcome after penile inversion vaginoplasty: a retrospective study of 475 transgender women. Plast Reconstr Surg. 2016;138(5):999-1007. https://doi.org/10.1097/PRS.0000000000002684

37. Hontscharuk R, Alba B, Hamidian Jahromi A, Schechter L. Penile inversion vaginoplasty outcomes: complications and satisfaction. Andrology. 2021;9(6):1732-43. doi: https://doi.org/10.1111/andr.13030

38. Papadopulos NA, Lellé JD, Zavlin D, Herschbach P, Henrich G, Kovacs L, et al. Quality of life and patient satisfaction following male-to-female sex reassignment surgery. J Sex Med. 2017;14(5):721-30. https://doi.org/10.1016/j.jsxm.2017.01.022

39. Klassen AF, Kaur M, Johnson N, Kreukels BP, McEvenue G, Morrison SD, et al. International phase I study protocol to develop a patient-reported outcome measure for adolescents and adults receiving gender-affirming treatments (the GENDER-Q). BMJ Open. 2018;8(10):e025435. https://doi.org/10.1136/bmjopen-2018-025435

40. van der Sluis WB, Schäfer T, Nijhuis TH, Bouman MB. Genital gender-affirming surgery for transgender women. Best Pract Res Clin Obstet Gynaecol. 2023;86:102297. https://doi.org/10.1016/j.bpobgyn.2022.102297

41. Dahlen S, Connolly D, Arif I, Junejo MH, Bewley S, Meads C. International clinical practice guidelines for gender minority/trans people: systematic review and quality assessment. BMJ Open. 2021;11(4):e048943. https://doi.org/10.1136/bmjopen-2021-048943

42. Zwickl S, Wong AF, Dowers E, Leemaqz SY, Bretherton I, Cook T, et al. Factors associated with suicide attempts among Australian transgender adults. BMC Psychiatry. 2021;21(1):81. https://doi.org/10.1186/s12888-021-03084-7

43. Dhejne C, Lichtenstein P, Boman M, Johansson AL, Långström N, Landén M. Long-term follow-up of transsexual persons undergoing sex reassignment surgery: cohort study in Sweden. PLoS One. 2011;6(2):e16885. https://doi.org/10.1371/journal.pone.0016885

44. Lee JL, Hirsh A, Radhakrishnan A, Jasuja GK, Taylor S, Dickinson S, Mineo J, Carnahan J, Weiner M. Mental health utilization among transgender veterans. JAMA Netw Open. 2025;8(1):e2454694. https://doi.org/10.1001/jamanetworkopen.2024.54694

45. Tordoff DM, Wanta JW, Collin A, Stepney C, Inwards-Breland DJ, Ahrens K. Mental health outcomes in transgender and nonbinary youths receiving gender-affirming care. JAMA Netw Open. 2022;5(2):e220978. https://doi.org/10.1001/jamanetworkopen.2022.0978

46. Summers NA, Huynh TT, Dunn RC, Cross SL, Fuchs CJ. Effects of gender-affirming hormone therapy on progression along the HIV care continuum in transgender women. Open Forum Infect Dis. 2021;8(9):ofab404. https://doi.org/10.1093/ofid/ofab404

47. Bränström R, Pachankis JE. Reduction in mental health treatment utilization among transgender individuals after gender-affirming surgeries: a total population study. Am J Psychiatry. 2020;177(8):727-34. https://doi.org/10.1176/appi.ajp.2019.19010080

48. Straub JJ, Paul KK, Bothwell LG, Deshazo SJ, Golovko G, Miller MS, et al. Risk of suicide and self-harm following gender-affirmation surgery. Cureus. 2024;16(4):e57472. https://doi.org/10.7759/cureus.57472

49. Achille, C., Taggart, T., Eaton, N. R., Osipoff, J., Tafuri, K., Lane, A., & Wilson, T. A. (2020). Longitudinal impact of gender-affirming endocrine intervention on the mental health and well-being of transgender youths: preliminary results. *International journal of pediatric endocrinology*, *2020*(1), 8.

50. Bos, P. A., Hermans, E. J., Ramsey, N. F., & van Honk, J. (2012). The neural mechanisms by which testosterone acts on interpersonal trust. *NeuroImage, 61*(3), 730–737. https://doi.org/10.1016/j.neuroimage.2012.04.002

51. Green, A. E., DeChants, J. P., Price, M. N., & Davis, C. K. (2022). Association of gender-affirming hormone therapy with depression, thoughts of suicide, and attempted suicide among transgender and nonbinary youth. *Journal of adolescent health*, *70*(4), 643-649.

52. Hughto, J. M., Gunn, H. A., Rood, B. A., & Pantalone, D. W. (2020). Social and medical gender affirmation experiences are inversely associated with mental health problems in a US non-probability sample of transgender adults. *Archives of sexual behavior*, *49*(7), 2635-2647.

53. Kiyar, M., Collet, S., T'Sjoen, G., & Mueller, S. C. (2020). Neuroscience in transgender people: An update. *Neuroforum, 26*(2), 85–92. https://doi.org/10.1515/nf-2019-0027

54. McEwen, B. S., & Milner, T. A. (2017). Understanding the broad influence of sex hormones and sex differences in the brain: Sex hormones affect the whole brain. *Journal of Neuroscience Research, 95*(1–2), 24–39. https://doi.org/10.1002/jnr.23809

55. Tack, L. J. W., Craen, M., Dhondt, K., Vanden Bossche, H., Laridaen, J., & Cools, M. (2016). Consecutive lynestrenol and cross-sex hormone treatment in biological female adolescents with gender dysphoria: A retrospective analysis. *Biology of Sex Differences, 7*(1), 14. https://doi.org/10.1186/s13293-016-0067-9

56. Tucker, R. P., Testa, R. J., Simpson, T. L., Shipherd, J. C., Blosnich, J. R., & Lehavot, K. (2018). Hormone therapy, gender affirmation surgery, and their association with recent suicidal ideation and depression symptoms in transgender veterans. *Psychological medicine*, *48*(14), 2329-2336.

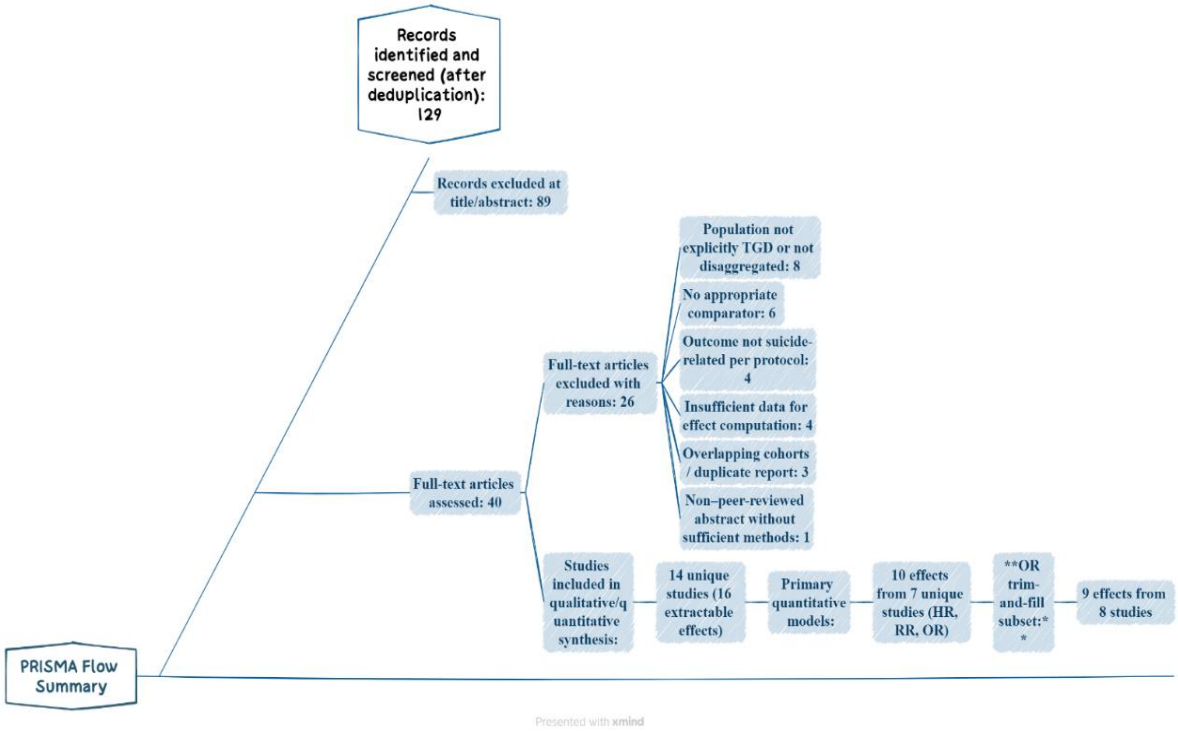

Figure 1. PRISMA Flow Diagram of Study Selection

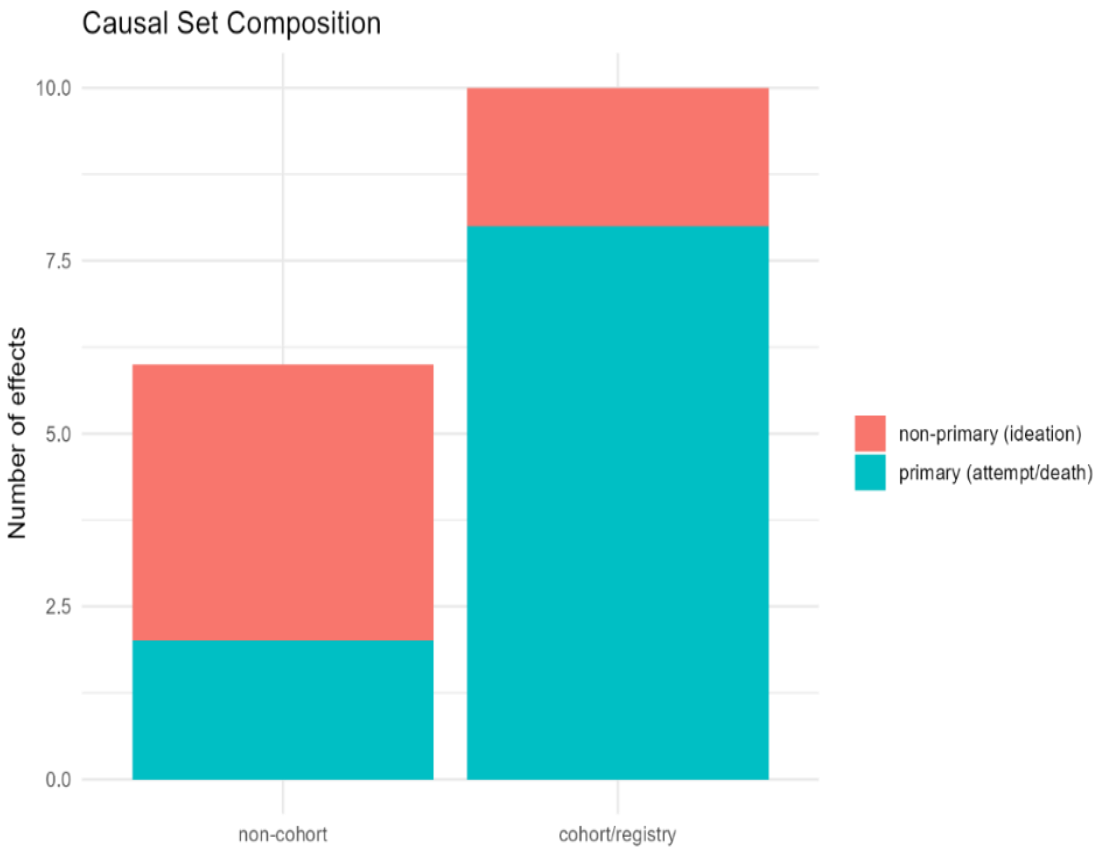

Figure 2. Distribution of cohort versus non-cohort study designs by primary causal inference framework. Study-level counts (n = 16) stratified by overarching design category. Cohort/registry studies (n = 10) predominantly employed registry linkage and prospective cohort methodologies, whereas non-cohort designs (n = 6) displayed greater methodological heterogeneity.

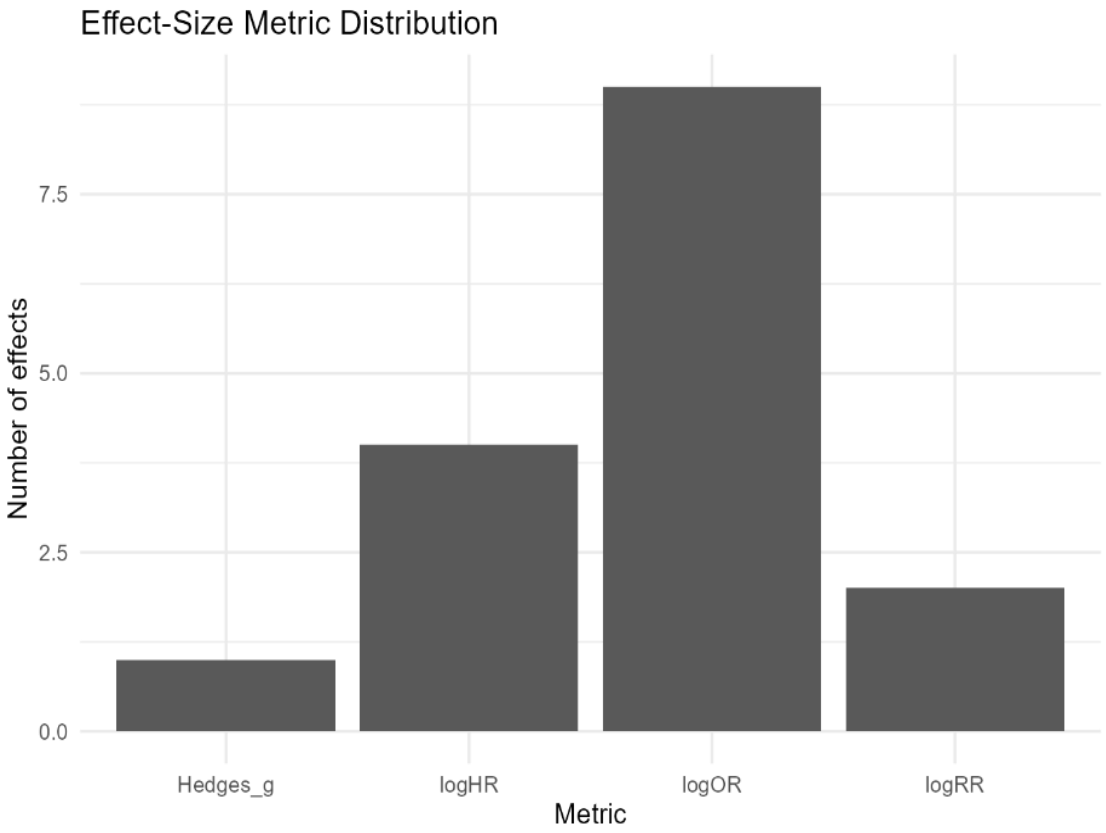

Figure 3. Frequency distribution of effect size metrics reported in included studies. Distribution of reported statistical metrics across studies (n = 16). Log odds ratios (logOR) were most frequent (n = 8), followed by log hazard ratios (logHR; n = 3), log relative risks (logRR; n = 2), and standardized mean differences (Hedges' g; n = 1).

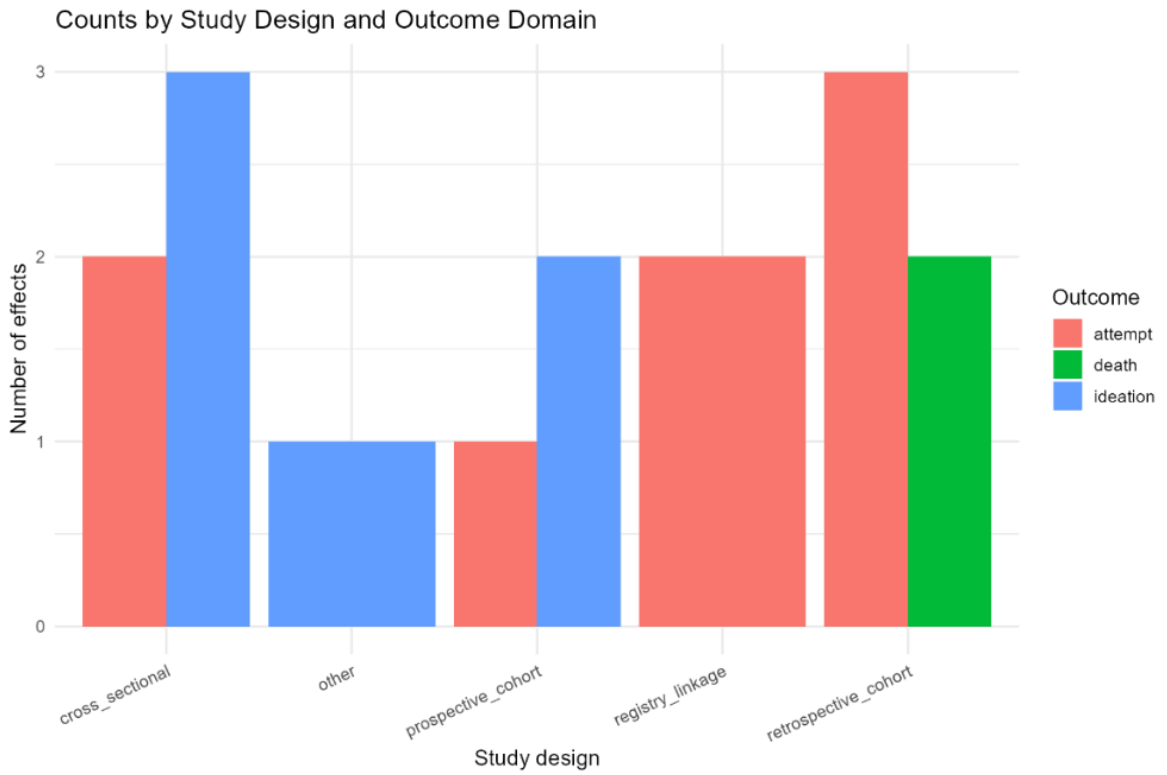

Figure 4. Cross-classification of study design types by primary suicide-related outcome category. Mapping of study designs (x-axis) against primary outcome domains (color-coded bars). Registry linkage studies were exclusively associated with suicide mortality outcomes, while other designs captured a broader range of suicidal behaviors, including ideation and attempts.

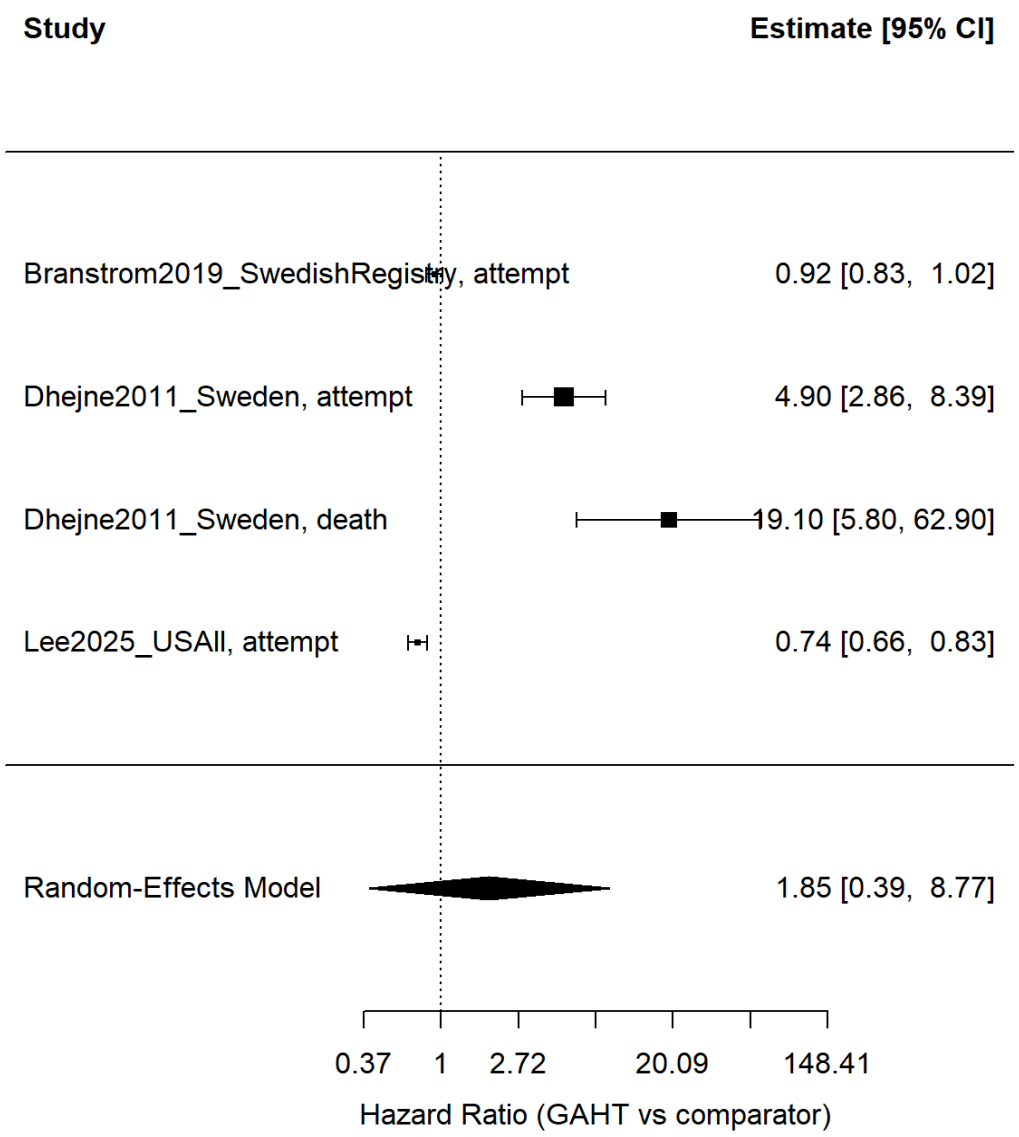

Figure 5. Random-effects meta-analysis of hazard ratios for suicide-related outcomes comparing GAHT recipients to non-recipients. Individual study estimates are shown as squares (size proportional to study weight) with 95% confidence intervals (horizontal lines); the pooled estimate is depicted as a diamond. HR values above 1 indicate increased hazard relative to the comparator group; values below 1 indicate reduced hazard. Between-study heterogeneity was substantial.

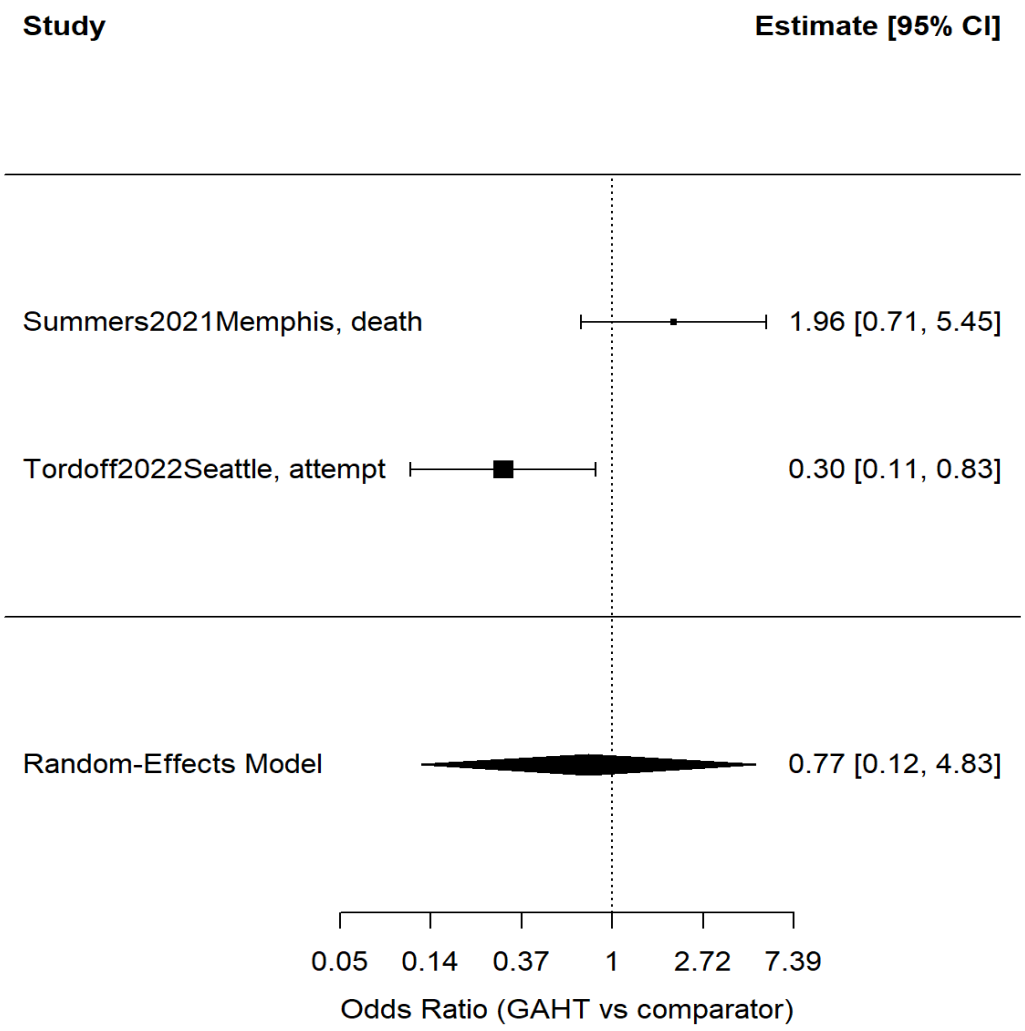

Figure 6. Random-effects meta-analysis of odds ratios for suicide-related outcomes comparing GAHT recipients to non-recipients. Individual studies and the pooled estimate are displayed as in Figure 5. OR values above 1 reflect higher odds of the outcome in GAHT recipients; values below 1 reflect lower odds. Considerable imprecision and variability across studies are evident.

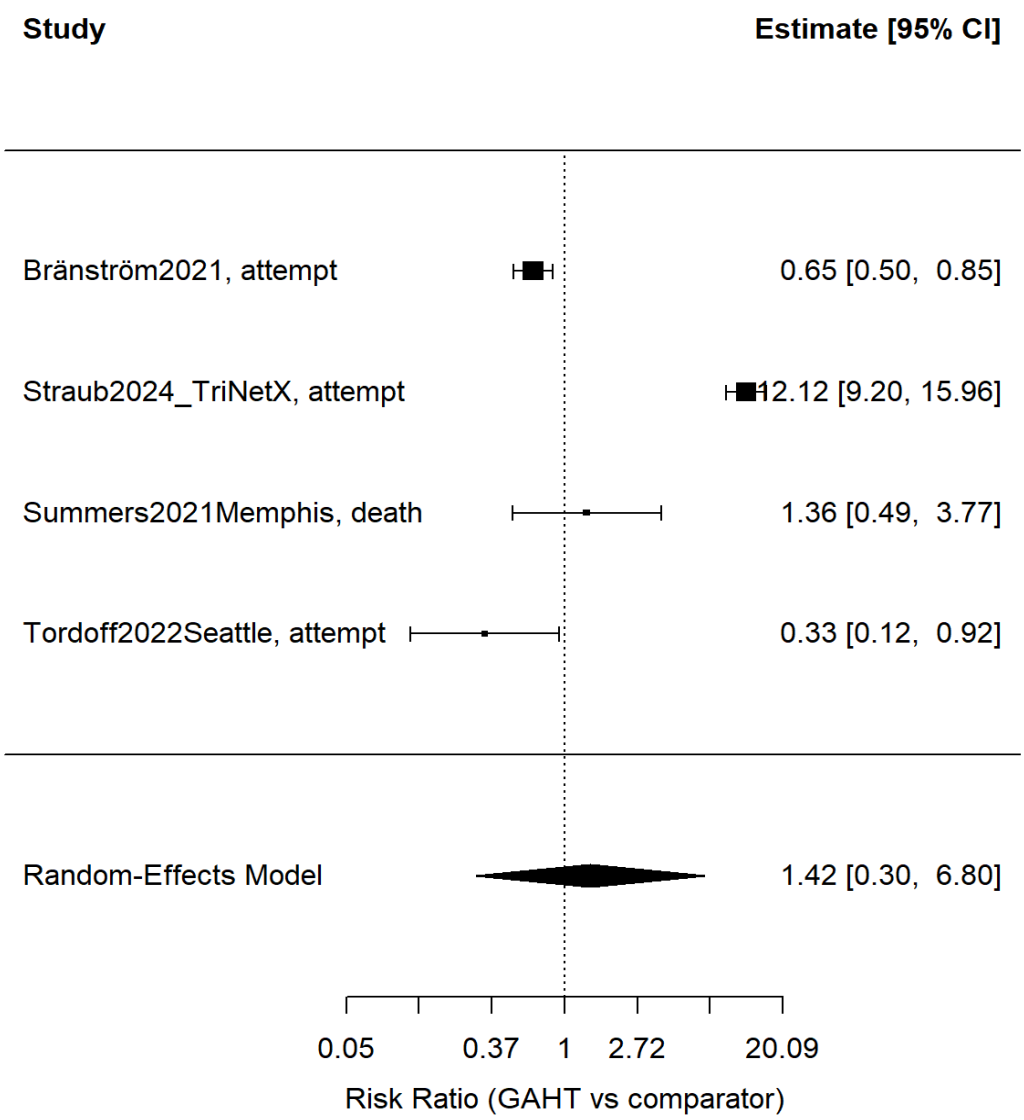

Figure 7. Random-effects meta-analysis of risk ratios for suicide-related outcomes comparing GAHT recipients to non-recipients. Individual study estimates and the pooled effect are presented as in Figures 5 and 6. RR values above 1 denote elevated risk in GAHT recipients; values below 1 denote reduced risk. Heterogeneity was high, and the prediction interval encompassed both protective and harmful associations.

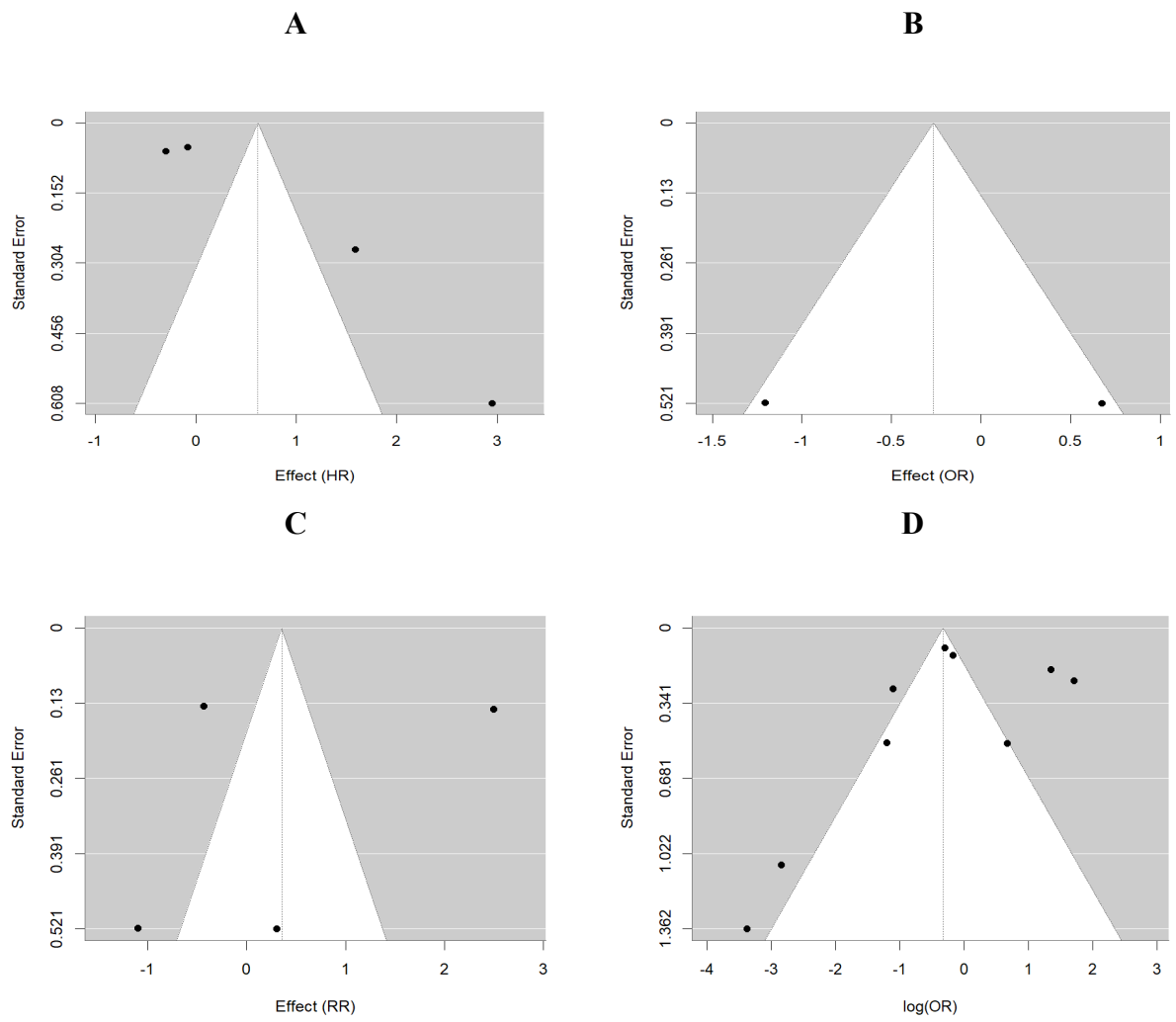

Figure 8. (A) Funnel plot of hazard ratio studies for suicide-related outcomes comparing GAHT recipients to non-recipients. Each point represents an individual study, plotted by effect size (x-axis) and standard error (y-axis). The distribution is moderately dispersed and largely symmetric around the pooled estimate, with two lower-precision, high-HR mortality studies appearing in the right tail. Visual inspection suggests no directional asymmetry indicative of small-study effects, consistent with the CR2 robust intercept test ($\beta = 0.618$, $p = 0.523$); **(B)** Funnel plot of odds ratio studies for suicide-related outcomes comparing GAHT recipients to non-recipients. The plot contains only two contributing studies, both located near the base of the funnel and equidistant from the pooled effect line. The extremely limited sample size precludes robust visual diagnosis of asymmetry, and the CR2 robust intercept test ($\beta = -0.266$, $p = 0.824$) did not indicate small-study effects. Interpretive caution is warranted due to the low k; **(C)** Funnel plot of risk ratio studies for suicide-related outcomes comparing GAHT recipients to non-recipients. Study points are symmetrically distributed around the pooled estimate, encompassing both protective and harmful associations. The absence of pronounced asymmetry is consistent with the CR2 robust intercept result ($\beta = 0.352$, $p = 0.690$). The dispersion primarily reflects genuine heterogeneity in study populations and endpoints rather than systematic publication bias; **(D)** Contour-enhanced funnel plot for odds ratio studies assessing suicide-related outcomes in GAHT recipients vs. non-recipients. Shaded regions denote conventional significance thresholds ($p < 0.10$, $p < 0.05$, $p < 0.01$). The filled black circle represents an imputed study from the trim-and-fill procedure.

Table 1. CR2-adjusted intercepts for small-study effect assessment across HR, OR, and RR meta-analytic models.

Metric	β (Intercept)	SE	t	df	p-value	Interpretation
HR	0.618	0.805	0.767	1.997	0.523	No statistical evidence of asymmetry
OR	–0.266	0.939	–0.283	1.000	0.824	No statistical evidence; interpretation limited by k = 2
RR	0.352	0.801	0.440	2.991	0.690	No statistical evidence of asymmetry

Table 2. Univariable meta-regression results for hazard ratio (HR) and risk ratio (RR) models, including variance components and pseudo-R^2_{Total}

Metric	Moderator	β (Estimate)	SE	p-value	Exp(β)	95% CI (Exp(β))	pseudo-R^2_{Total}	Interpretation
HR	Outcome = Death	2.010	1.109	0.070	7.468	0.849, 65.676	0.485	Suggestive but non-significant hazard elevation for mortality outcomes
HR	Country = Other	-1.381	1.961	0.481	0.251	0.005, 11.719	0.000	No detectable geographic effect
HR	GAHT + Other Intervention	2.072	0.327	<0.001	7.941	4.190, 15.048	0.982	Strong, significant hazard elevation; high variance explained
RR	Outcome = Death	-0.048	2.254	0.983	0.952	0.012, 78.963	0.000	No effect detected
RR	Timeframe = Other	-1.904	1.855	0.305	0.149	0.004, 5.648	0.023	No significant moderation
RR	Country = Other	1.429	1.671	0.393	4.174	0.157, 110.358	0.000	No significant moderation
RR	GAHT + Other Intervention	1.429	1.671	0.393	4.174	0.157, 110.358	0.000	No significant moderation

Table 3. Trim-and-fill adjusted odds ratio estimates for suicide-related outcomes comparing GAHT recipients to non-recipients.

Model	k (Observed)	k (Imputed)	OR (Unadjusted)	95% CI (Unadj.)	OR (Adjusted)	95% CI (Adj.)
Random-effects OR	8	1	0.77	0.12 – 4.83	0.91	0.15 – 5.31