Do Psychosocial Interventions Fade Out? Evidence from 1,503 Meta-Analyses

Duncan Webb*

March 26, 2026

Abstract

Cost-effectiveness analyses of social programs depend critically on how long treatment effects last, yet most interventions lack long-run follow-up data. We provide the first large-scale empirical prior on this parameter. Drawing on 1,503 meta-analyses from Cochrane systematic reviews of psychosocial interventions, we estimate treatment effect trajectories over time using Bayesian hierarchical models. Fadeout is small: only 5.5% of analyses show detectable time trends, and the average effect trend is approximately 0.01 SD per year in magnitude, statistically indistinguishable from zero—economically negligible relative to mean initial effects of 0.34 SD. These findings survive power analyses, selection corrections, and pessimistic bounds on missing long-run data. In the absence of domain-specific evidence to the contrary, our results suggest that policymakers and funders can reasonably assume meaningful persistence within the first one to two years—rather than the rapid fadeout that many cost-effectiveness models implicitly assume.

JEL Codes: *C11, C83, I10, I12, J24*

Keywords: *fadeout, treatment effect persistence, meta-analysis, psychosocial interventions, cost-effectiveness*

*Nova School of Business and Economics. Email: duncan.webb@novasbe.pt. I am grateful to Ryan Briggs, Anna Dreber, Michael Gechter, Adam Salisbury, and Eva Vivalt for valuable comments and feedback, and to seminar participants at Nova SBE. Joseph Levine provided excellent research assistance. This work was generously supported by Coefficient Giving. I acknowledge support from Fundação para a Ciência e a Tecnologia (UID/00124/2025, UID/PRR/124/2025, Nova School of Business and Economics).

1 Introduction

A policymaker considering whether to fund a psychosocial intervention—psychotherapy, early childhood education, a behavioral health program—faces a decision that hinges on one largely unknown parameter: how long will the benefits last? If treatment effects persist, even a modestly effective program can be highly cost-effective; if they fade quickly, the same investment may yield little lasting value. The assumed rate of fadeout can change the net present value of an intervention by a factor of two or more. Yet for most interventions, long-run follow-up data simply do not exist. Policymakers must therefore rely on assumptions about persistence—assumptions that, until now, have been largely ungrounded in systematic evidence.

A substantial literature documents fadeout—the tendency of treatment effects to diminish over time. In economics, fadeout is central to cost-effectiveness analysis: the present value of an intervention depends critically on how long benefits persist, yet this parameter is rarely measured directly (D. Bailey, Duncan, Odgers, & Yu, 2017). Meta-analyses of IQ-raising interventions estimate fadeout rates of approximately 0.02 standard deviations per month (Protzko, 2015). In development economics, unconditional cash transfers and entrepreneurial grants frequently show declining effects over time (Bouguen, Huang, Kremer, & Miguel, 2019). Models of skill formation predict that early investments may depreciate without complementary later investments, providing a theoretical mechanism for fadeout (Cunha & Heckman, 2007). These patterns have led some to question whether social interventions produce lasting benefits at all (Cascio & Staiger, 2012).

Yet fadeout is far from universal. Other studies have documented lasting effects. The Jamaica early stimulation trial found that psychosocial intervention in infancy produced 25% higher earnings two decades later (Gertler et al., 2014). The Perry Preschool and Abecedarian programs showed lasting effects on adult health, crime, and economic outcomes (Conti, Heckman, & Pinto, 2016). Head Start itself shows persistent effects on adult outcomes despite fading test scores (Almond, Currie, & Duque, 2018). Cognitive behavioral therapy reduced crime and violence over ten years in Liberia (Blattman, Chaskel, Jamison, & Sheridan, 2023), and mass deworming in Kenya produced sustained earnings gains two decades later (Baird, Hicks, Kremer, & Miguel, 2016). In clinical psychology, meta-analyses of cognitive behavioral therapy find that 70–80% of post-treatment gains persist at follow-up (Rith-Najarian, Mesri, Lane, Engstrom, & Chavira, 2019). Child health, nutrition, and cognitive stimulation interventions tend to produce durable benefits (Bouguen et al., 2019).

This heterogeneous evidence leaves policymakers and funders without an answer to a question they face routinely: when long-run follow-up data are unavailable for a specific intervention, what should they assume about persistence? By psychosocial interventions, we mean treatments that operate through psychological, behavioral, or social mechanisms rather than pharmacological ones—including psychotherapy, cognitive behavioral therapy, patient education, exercise programs, social support, and care management. For this broad class, domain-specific long-run evidence is the exception, not the rule. Cost-effectiveness analyses must therefore rely on assumptions about fadeout, yet the scenarios these assumptions generate are radically different. Under rapid fadeout, benefits

are exhausted within months and the intervention is a temporary palliative; under persistence, the same program represents a durable investment in human capital. An informed prior—grounded in evidence about typical fadeout rates across similar interventions—would allow policymakers to distinguish between these scenarios rather than guess. No such prior currently exists. Generating one requires synthesizing evidence at a scale beyond what any individual study or narrative review can achieve.

We provide the first large-scale quantitative estimate of treatment effect persistence across psychosocial interventions. Our analysis draws on the Cochrane Database of Systematic Reviews, the gold standard for evidence synthesis in health and social interventions. From this database, we extract 1,503 distinct "analyses"—each pooling effect sizes for a single intervention-outcome pair measured at different follow-up periods. This scope encompasses interventions ranging from psychotherapy and patient education to exercise programs and social support, with outcomes spanning mental health, physical health, behavior change, and functional status. We focus on psychosocial interventions because the Cochrane Database provides a uniquely structured corpus for this class—thousands of comparable analyses measuring outcomes at multiple follow-up periods—and because these interventions share core mechanisms with the social programs economists study in development and education settings: behavior change, skill acquisition, and social support. In ongoing work, we extend this framework to interventions in development economics, including cash transfers, job training, and early childhood programs.

Our methodological contribution is to explicitly model treatment effect dynamics using hierarchical models—to our knowledge, the first application of this approach at scale. Standard meta-analysis pools effects measured at different follow-up periods without modeling how effects evolve over time; we instead parameterize a linear time trend within each analysis, separating between-study heterogeneity from temporal dynamics. We then aggregate these analysis-level slope estimates to characterize the distribution of fadeout rates across interventions. This framework yields three key parameters: the average fadeout rate across all analyses, the variance of fadeout rates (capturing heterogeneity), and the correlation between initial effect size and fadeout rate—parameters not previously estimated at this scale.

Our results indicate that fadeout is small. Only 83 of 1,503 analyses (5.5%) show statistically detectable time trends. The average fadeout rate is approximately 0.0008 SD per month—equivalent to roughly 0.01 SD per year—with the 95% credible interval spanning zero. For an intervention with an initial effect of 0.34 SD (the sample mean), linearly extrapolating this rate implies negligible cumulative fadeout—though we stress that the data are concentrated at shorter horizons (median follow-up of 6.1 months) and projections beyond observed horizons carry substantial uncertainty. We also find a negative correlation between initial effects and fadeout rates ($\rho \approx -0.35$): larger initial effects tend to fade faster, consistent with mean reversion. This finding—that the most impressive short-run results are least likely to persist—has direct implications for how policymakers should forecast the long-run returns of social programs.

These results could reflect low statistical power rather than genuine persistence—perhaps true

fadeout exists but our analyses lack the precision to detect it. We address this concern through simulation-based power analysis. Among 163 high-quality analyses with favorable statistical properties, we find that 60–80% would achieve 80% power to detect fadeout of 0.03 SD per month. Yet the empirical detection rate shows no relationship with simulated power: better-powered analyses are no more likely to find fadeout than worse-powered ones. This flat pattern is inconsistent with the hypothesis that low power explains our null results.

Selection poses a more fundamental threat. The key concern is that follow-up timing is itself endogenous: researchers who expect an intervention's effects to fade quickly have an incentive to measure outcomes sooner, before effects disappear; those who expect persistence may invest in longer follow-up to document durable benefits. If this selection operates systematically, our data will over-represent persistence and under-represent fadeout—we simply never observe the long-run outcomes of interventions expected to fade. This is not a bias we can eliminate, but we take it seriously through four complementary approaches.

First, we test for selection on short-run results directly. If researchers extend follow-up only when early results look promising, studies with long-run data should have systematically larger short-run effects. Within-analysis comparisons reject this prediction: studies with long-run follow-up show short-run effects that are -0.009 SD *lower* than short-run-only studies (SE $= 0.015$), and the probability of long-run follow-up is flat across deciles of residualized short-run effects.

Second, we use inverse probability weighting to correct for selection on observable characteristics—sample size, precision, intervention type. If analyses with certain features are both more likely to include long-run follow-up and more likely to show persistence, our estimates would be biased. IPW correction yields fadeout estimates of -0.0007 to -0.0025 SD/month, *slightly larger* than uncorrected estimates. Observable selection, if anything, works against persistence: analyses most likely to include long-run data show less decline, so correcting for this increases estimated fadeout.

Third, and most importantly, we use exogenous variation in NIH disease-area funding as an instrument for follow-up duration. The identifying variation comes from year-to-year fluctuations in categorical NIH spending interacted with the disease-area composition of each Cochrane review, producing a Bartik-style shift-share instrument. A control function approach reveals that selection understates fadeout by roughly 0.030 SD per year, yielding a corrected estimate of approximately -0.099 SD per year. The direction of selection confirms the intuition that researchers invest in longer follow-up when they expect persistence—but the correction is modest, and the qualitative conclusion of meaningful persistence holds.

Fourth, as a worst-case exercise, we construct pessimistic bounds that assume all missing long-run data would have shown complete fadeout to zero. For every study lacking observations beyond 24 months, we augment the data with synthetic zero-effect observations at 36 months and re-estimate the model. Under this bound, annual fadeout increases to 0.09 SD per year. An even more extreme bound, assuming *all* studies would show complete fadeout shortly after their last observation, yields 0.25 SD per year—roughly one-quarter to three-quarters of the initial effect fading within the first

year. We do not dismiss these bounds: the endogeneity of follow-up timing is a genuine limitation. But the instrumental variable evidence suggests the truth lies closer to our baseline estimates than to these extremes.

Our findings have implications for both research and policy. For researchers studying treatment effect dynamics, the low rate of detectable fadeout (5.5%) suggests that static meta-analyses—which ignore when outcomes are measured—may provide adequate summaries for most intervention literatures. For policymakers and funders, our results speak directly to the decision problem motivating this paper: what to assume about persistence when intervention-specific long-run data are unavailable. The two scenarios that policymakers must distinguish between—rapid fadeout versus meaningful persistence—imply very different cost-effectiveness conclusions. Consider an intervention costing $1,000 per person with an initial effect of 0.34 SD. Under a pessimistic assumption of 0.05 SD per month fadeout, benefits are exhausted within seven months. Under our estimate of roughly 0.01 SD per year, the model projects that most of the initial effect persists well beyond the first year—though this extrapolation rests on limited data at longer horizons, since only 5,154 observations (across 1,741 analyses) have follow-up exceeding 12 months. In the absence of domain-specific evidence to the contrary, our results provide an empirically grounded prior: for a typical psychosocial intervention, policymakers can reasonably assume meaningful persistence within the first one to two years, rather than defaulting to the rapid fadeout that many cost-effectiveness models implicitly assume.

Related Literature. Empirical work on persistence has proceeded domain by domain. Meta-analyses examine durability of cognitive behavioral therapy (Cuijpers et al., 2013), mindfulness-based relapse prevention (Kuyken et al., 2016), digital mental health interventions (G. Andrews et al., 2018), parenting programs (Barlow, Bergman, Kornør, Wei, & Bennett, 2016), and early childhood education (D. Bailey et al., 2017; D. H. Bailey, Duncan, Cunha, Foorman, & Yeager, 2020). Each review reaches conclusions specific to its intervention type and outcome. The closest work in economics is Bhat et al. (2022), who study long-run effects of psychotherapy across two trials in India; our contribution is to estimate fadeout rates systematically across the entire class of psychosocial interventions. What remains unknown is whether these domain-specific patterns aggregate to a general tendency toward persistence or fadeout across psychosocial interventions as a class.

Methodologically, we depart from standard meta-analysis, which pools effect sizes from studies with different follow-up periods without modeling how effects evolve over time (DerSimonian & Laird, 1986). We explicitly parameterize fadeout dynamics within a hierarchical framework, following Meager (2019) in using Bayesian hierarchical models to aggregate heterogeneous treatment effects across studies. This builds on the umbrella review approach that synthesizes across systematic reviews (Ioannidis, 2009). The heterogeneity we document connects to Vivalt's (2020) finding that treatment effects vary substantially across contexts; our contribution is to separate temporal variation (fadeout) from cross-study heterogeneity. This yields the first cross-domain quantitative estimate of treatment effect persistence and the first estimate of the correlation between initial effect

magnitude and subsequent fadeout—a parameter that speaks directly to whether larger effects are more or less durable.

2 Data and Methods

2.1 Data

We extract effect size estimates from the Cochrane Database of Systematic Reviews, focusing on psychosocial interventions—behavioral, psychological, and social support treatments. These include psychotherapy (e.g., cognitive behavioral therapy, problem-solving therapy), patient education programs, care management approaches, physical exercise interventions, and support programs (Table 1). We classify reviews as psychosocial using a large language model applied to review abstracts (Appendix C). Figure A1 summarizes the sample construction process.

Table 1 summarizes the data. Panel A describes the two samples used in our analysis. For the analysis-level model (Section 3.1), we restrict to analyses with at least three contributing studies and sufficient temporal variation (interquartile range of follow-up times ≥ 3 months) to permit reliable estimation of effect dynamics in individual analyses. For the aggregate model (Section 3.2), we include all analyses with any temporal variation, maximizing statistical power to detect average fadeout patterns. Panel B characterizes the intervention types, identified by clustering PICO intervention terms from Cochrane reviews using text embeddings: psychotherapy dominates (254 reviews, spanning cognitive behavioral, family, and movement therapies) but the sample also includes health education, care planning, workplace and technology-based interventions, social support programs, substance use cessation therapies, exercise interventions, skills training, and breastfeeding support. Outcomes are similarly diverse, spanning anxiety and depression severity, functional outcomes, physical health, substance use, quality of life, and healthcare utilization, among others (Table B4).

For each observation, we record the effect size, its standard error, and the follow-up period—defined as time from treatment completion to outcome measurement. We extract follow-up periods from outcome labels using GPT-5, with validation described in Appendix D; Figure D7 shows the distribution. Effect sizes fall into two categories: *difference measures*—standardized mean differences (SMD) and mean differences (MD)—and *ratio measures*—odds ratios (OR) and risk ratios (RR). To enable pooling across all effect types, we convert all effects to a common scale of standard deviation (SD) units. For difference measures, SMDs are used directly while MDs are divided by the pooled standard deviation. For ratio measures, we use the probit transformation: $d = \Phi^{-1}(p_e) - \Phi^{-1}(p_c)$, where p_e and p_c are observed event rates in treatment and control groups, and Φ^{-1} is the inverse standard normal CDF (Glass, McGaw, & Smith, 1981).[1]

Because treatment effect directions are inherently arbitrary—an intervention might reduce symptoms (negative effect) or increase well-being (positive effect)—we standardize directions within

[1]This "probit d" approach directly computes the standardized difference on the latent normal scale, avoiding distributional assumptions required by log-odds transformations. When event counts are unavailable (18% of ratio measures), we fall back to the Chinn (2000) log-odds transformation. See Appendix E for details.

Table 1: *Sample Descriptives and Intervention Types*

Panel A: Sample Overview

	Analysis-Level (Section 3.1)	Aggregate (Section 3.2)
Inclusion criteria	IQR $\geq$ 3 mo, $\geq$ 3 studies	Any time variation
Analyses	1,503	4,738
Observations	16,007	35,404
Cochrane reviews	248	480
Unique studies	3,693	6,759
Median follow-up (months)	6.1	3.0
IQR follow-up (months)	1.0–12.2	0.0–7.6

Panel B: Intervention Types

Category	Reviews	Representative Interventions	Example Review
Psychotherapy	254	Talking therapy, Movement Therapy, Psychotherapy	*"Psychological interventions for depression in adolescent and adult congenital heart disease"*
Multi-component Interventions	138	Psychosocial Intervention Strategy, Multiple Component Intervention, Preventative intervention	*"Psychosocial interventions for conversion and dissociative disorders in adults"*
Health Education	136	Patient Education, Health education, Diet Education	*"Patient and lay carer education for preventing pressure ulceration in at-risk populations"*
Care Planning	118	Personalized Care Planing, Advance Care Planning, Patient-centered care	*"Personalised care planning for adults with chronic or long-term health conditions"*
Workplace & Technology	116	Work Environment, Work tasks, Telephone	*"Workplace interventions for increasing standing or walking for decreasing musculoskeletal symptoms in sedentary workers"*
Social Support	89	Caregiver Support, Peer support, Supportive Counseling	*"Remotely delivered information, training and support for informal caregivers of people with dementia"*
Substance Use Cessation	84	Substance use cessation therapy, Smoking Cessation Drug Therapy, Smoking Cessation Therapy	*"Interventions for tobacco use cessation in people living with HIV"*
Exercise & Physical Activity	72	Exercise therapy, Physical Exercises, Physical activity	*"Effects of additional exercise therapy after a successful vascular intervention for people with symptomatic peripheral arterial disease"*
Skills Training	63	Functional Skills Training, Habilitation Training, Skills Training	*"Oral hygiene interventions for people with intellectual disabilities"*
Breastfeeding Support	38	Breastfeeding Education, Breastfeeding Support Education, Breastfeeding Support	*"Treatments for breast engorgement during lactation"*

Notes: Intervention terms from PICO descriptions in Cochrane reviews, clustered using k-means on text embeddings. "Reviews" indicates the number of psychosocial reviews (out of the full corpus, before applying the sample restrictions in Panel A) containing intervention terms in each cluster. Reviews may appear in multiple clusters if their PICO descriptions span categories. Representative interventions are the terms closest to each cluster centroid.

each analysis so that the inverse-variance weighted mean effect is positive. This is purely a sign convention: it does not assume the intervention is beneficial, only that the main effect is coded as positive so that "fadeout" consistently means effects moving toward zero over time. Specifically, if the weighted mean effect within an analysis is negative, we flip the sign of all effect sizes in that analysis (Appendix F.5). If an intervention is in fact harmful (i.e., the positive-coded effect reflects harm), fadeout still means the effect shrinks toward zero—harm dissipates—which is correctly captured as a negative slope.

2.2 Defining an Analysis

Our approach estimates a separate fadeout rate for each distinct research question, then aggregates these estimates. We therefore need to define what constitutes a single "analysis"—the unit on which we run each meta-analysis. An *analysis* pools effect sizes for the same intervention-outcome pair measured at different follow-up periods. Specifically, each analysis combines estimates that share: (i) the same outcome measure (e.g., depression symptoms measured via the Beck Depression Inventory), and (ii) the same comparison of interventions (e.g., cognitive behavioral therapy versus waitlist control). Details on edge cases are in Appendix G. This yields 1,503 distinct analyses.

2.3 Analysis-Level Model

For each analysis j, we estimate a Bayesian hierarchical model with a linear time trend:

$$Y_{ijt} \sim N(\theta_{ij} + \beta_j \cdot t_{ijt}, \sigma^2_{ijt}) \tag{1}$$

$$\theta_{ij} \sim N(\alpha_j, \tau^2_j) \tag{2}$$

where Y_{ijt} is the effect size for study i in analysis j at follow-up time t (measured in months), expressed in SD units as described above; σ_{ijt} is the reported standard error (treated as known); θ_{ij} is a study-specific intercept; α_j is the population mean intercept; β_j is the fadeout rate (common across studies within the analysis); and τ_j captures between-study heterogeneity.

We place Normal priors on α_j and β_j, and a half-Cauchy prior on τ_j, with hyperparameters calibrated to the SD scale (Appendix F). Models are estimated via Markov chain Monte Carlo using Stan.

The key output is the posterior distribution of the slope β_j. We say an analysis shows "evidence of dynamics" if the posterior probability of direction exceeds 0.95—that is, if $\max\{P(\beta_j > 0 \mid \text{data}), P(\beta_j < 0 \mid \text{data})\} > 0.95$, or equivalently, if the 90% credible interval excludes zero. A negative β_j indicates fadeout (effects declining over time); a positive β_j indicates effect growth.

Identification. Estimating β_j requires distinguishing between-study heterogeneity (τ_j) from temporal dynamics: if Study A finds a larger effect than Study B, is that because Study A is genuinely more effective, or because it measured outcomes earlier? The study-specific intercepts θ_{ij} absorb cross-study differences in effect magnitude, allowing β_j to be identified from variation in follow-up time. *Panel studies*—studies that measure outcomes at multiple follow-up periods—are particularly

valuable because they provide within-study variation in t, allowing clean identification of β_j while controlling for study-level differences. Of the 3,693 studies in our sample, 1,159 (31.4 %) are panel studies. Critically, 45.4 % of analyses contain at least one panel study. Even analyses without panel studies contribute to identification through between-study variation in follow-up timing, though with less precision.

2.4 Aggregating Across Analyses

To estimate the overall average fadeout rate, we fit a hierarchical model pooling all observations across all analyses:

$$Y_{ijt} \sim N(\theta_i + \alpha_j + \beta_j \cdot t_{ijt}, \sigma^2_{ijt}) \tag{3}$$

$$\theta_i \sim N(0, \tau^2_{\text{study}}) \tag{4}$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma \right) \tag{5}$$

where σ_{ijt} is the reported standard error (treated as known), θ_i is a study-specific random effect, and Σ allows correlation between intercepts and slopes:

$$\Sigma = \begin{pmatrix} \tau^2_\alpha & \rho\,\tau_\alpha\tau_\beta \\ \rho\,\tau_\alpha\tau_\beta & \tau^2_\beta \end{pmatrix} \tag{6}$$

The study effect θ_i is crucial because the same study often appears in multiple analyses—measuring different outcomes on the same participants. Approximately 80% of studies contribute to two or more analyses. Without accounting for this dependence, we would treat correlated effect sizes as independent, underestimating standard errors for our aggregate estimates.[2]

The key parameters are μ_α (average initial effect across analyses), μ_β (average fadeout rate), τ_{study} (between-study heterogeneity), τ_α and τ_β (between-analysis heterogeneity in intercepts and slopes), and ρ (the correlation between intercepts and slopes). A negative ρ would indicate that larger initial effects tend to fade faster; a positive ρ would indicate they persist longer. Because all effect sizes have been converted to SD units (Section 2), we estimate a single unified model pooling all effect types.

3 Main Results

3.1 Analysis-Level Results

Figure 1 presents our central finding: most analyses show no detectable time trend in treatment effects. Panel (a) displays the distribution of estimated fadeout rates ($\hat{\beta}_j$) across 1,503 analyses. The distribution is sharply peaked around zero. Only 83 analyses (5.5%) show evidence of dynamics,

[2]The model assumes that, conditional on the linear time trend, residuals are uncorrelated across time points. The trend $\beta_j \cdot t$ captures how effects evolve over time; any remaining temporal dependence in the errors would reflect correlation beyond what the trend explains.

meaning their 90% credible interval excludes zero. Using the more stringent 95% credible interval, only 37 analyses (2.5%) show detectable dynamics. Among those detected at the 90% level, 49 show fadeout (declining effects over time) and 34 show growth (increasing effects).

The estimated slopes cluster tightly near zero. Across all 1,503 analyses, the median slope is approximately -0.002 SD per month, with an interquartile range of −0.02 to 0.01. Even at the tails of the distribution, slopes remain small: annual effect changes rarely exceed ±0.5 SD.

Panel (b) presents an alternative way to assess how often dynamics matter, by asking: does accounting for time trends change conclusions relative to ignoring them? To answer this, we compare two models for each analysis. The *static model* is a standard random-effects meta-analysis that pools all effect sizes without accounting for when outcomes were measured. The *dynamic model* is our linear trend specification, which predicts different effects at different follow-up times. We then compare the static model's pooled estimate to the dynamic model's predictions at two time points: the *short-run effect* (the earliest follow-up time observed in that analysis) and the *long-run effect* (the latest follow-up time observed). We say conclusions "differ" if either (i) the 95% credible intervals do not overlap, or (ii) one estimate is statistically significant while the other is not.

At the short-run horizon, 14.8% of analyses yield different conclusions under the dynamic versus static model. At the long-run horizon, this rises to 25.7%, because even small slopes compound over time. Overall, 32.8% of analyses show some change in conclusions when accounting for dynamics.

The low rate of detectable dynamics (5.5%) admits two interpretations: either treatment effects are genuinely persistent, or we lack statistical power to detect fadeout. Figure B3 suggests both may be at play. The figure plots each analysis's slope estimate against its credible interval width. Few analyses have large absolute slopes, consistent with genuinely small effects. But many analyses also have wide credible intervals, indicating limited power to detect even moderate fadeout. We explore power more systematically in Section 5.

3.2 Aggregate Fadeout Estimates

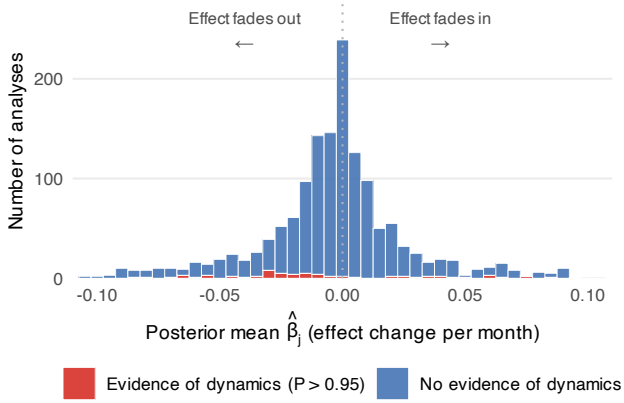
While individual analyses often lack statistical power to detect fadeout, pooling across all analyses using the crossed random effects model (Section 2.4) provides a precise aggregate estimate. Figure 2 presents the estimated average effect trajectory, with parameter estimates reported in Panel (b). Because the aggregate model does not require precise estimation within each analysis, we include all analyses with any temporal variation—a substantially larger sample than the analysis-level model (4,738 analyses comprising 35,404 observations).

The average fadeout rate is $\hat{\mu}_\beta = 0.0008$ SD per month (95% CI: $[-0.003, 0.004]$), with the credible interval including zero and the point estimate slightly positive. The posterior probability that the true average slope is negative is only 33%—the data favor slight effect *growth* rather than fadeout, though neither direction is statistically distinguishable from zero. In annual terms, this corresponds to approximately 0.01 SD per year. The mean initial effect is $\hat{\mu}_\alpha = 0.34$ SD (95% CI: $[0.33, 0.36]$).

These aggregate estimates indicate that average fadeout is small. The 95% credible interval for the

Figure 1: *Distribution of Fadeout Rates and Evidence of Effect Dynamics*

(a) Distribution of $\hat{\beta}_j$



(b) Summary Statistics

Metric	N	%
90% CI excludes zero	83	5.5%
Short-run ≠ static	223	14.8%
Long-run ≠ static	387	25.7%
Any conclusion change	493	32.8%
Total analyses	1503	100%

Notes: Panel (a) shows the distribution of estimated fadeout rates ($\hat{\beta}_j$, the posterior mean slope for each analysis j) across 1,503 analyses. Each bar represents one analysis; red bars indicate analyses with "evidence of dynamics" (90% credible interval excludes zero), blue bars indicate no detectable trend. The dashed vertical line marks zero (no fadeout). Panel (b) summarizes how often accounting for effect dynamics changes conclusions relative to standard (static) meta-analysis. "90% CI excludes zero": analyses where the slope's credible interval does not contain zero, indicating statistically detectable fadeout or growth. "Short-run ≠ static": analyses where the dynamic model's prediction at treatment completion ($t = 0$) differs meaningfully from the static pooled estimate. "Long-run ≠ static": analyses where the dynamic model's prediction at 24 months post-treatment differs from the static estimate. "Any conclusion change": analyses meeting any of the above criteria.

mean slope spans zero, and the point estimate is economically negligible—roughly 0.01 SD per year relative to mean initial effects of 0.34 SD. An important caveat is that the data are concentrated at shorter horizons: 13,548 observations (across 3,211 analyses) have follow-up exceeding 6 months, 5,154 observations (1,741 analyses) exceed 12 months, and only 1,047 observations (572 analyses) exceed 24 months. The aggregate estimate is therefore most informative about fadeout within the first one to two years; posterior predictive intervals widen considerably at longer horizons where data are sparse.

Larger effects fade faster. The model estimates a substantial negative correlation ($\rho = -0.35$, 95% CI: $[-0.38, -0.31]$) between initial effect sizes and fadeout rates.[3] This is one of our most precisely estimated and policy-relevant findings. Analyses with larger initial effects tend to fade faster, while those with smaller initial effects are more likely to persist or grow. The implication for policymakers is that interventions producing the most dramatic short-run improvements may require follow-up support or booster sessions to sustain gains. Conversely, interventions with modest but reliable initial effects may be more durable investments. One interpretation is mean reversion: unusually large observed effects may partly reflect favorable circumstances or measurement timing, and thus regress toward more modest long-run values. This pattern is consistent with findings in the broader program evaluation literature that treatment effects vary substantially across contexts (Vivalt, 2020); our contribution is to show that this heterogeneity has a systematic temporal

[3]This correlation is not an artifact of direction standardization: because flipping reverses both the intercept and slope simultaneously, the *relationship* between intercept magnitude and slope is preserved on the original scale.

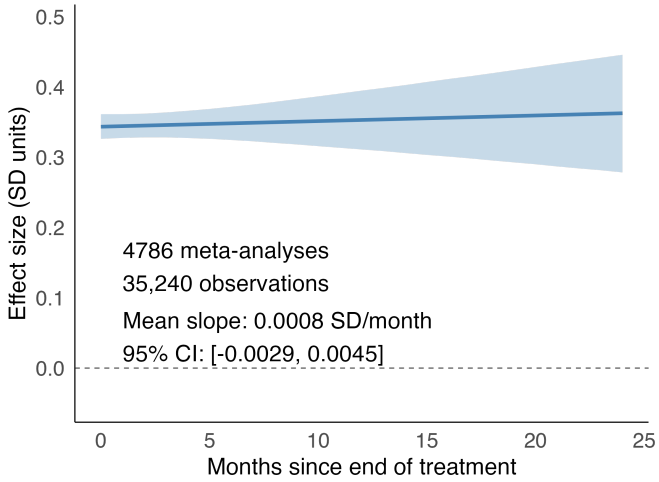
structure.

The heterogeneity in slopes ($\tau_\beta = 0.1$) indicates substantial variation across analyses—some show fadeout, others show growth—even though the average is near zero.[4]

These estimates rest on the assumption that follow-up timing is unrelated to expected effect trajectories—that researchers do not systematically choose to measure outcomes at times when effects are expected to be larger or smaller. We examine the plausibility of this assumption and its implications in Section 4.

Figure 2: *Average Fadeout Curve and Model Parameters*

(a) Average Effect Trajectory



(b) Model Parameters

Parameter		Est. [95% CI]
μ_α	Mean initial effect (SD)	0.344 [0.326, 0.361]
μ_β	Mean fadeout rate (SD/mo)	0.0008 [-0.0029, 0.0045]
ρ	Intercept-slope corr.	-0.348 [-0.382, -0.313]
τ_α	Between-analysis SD (int.)	0.264 [0.255, 0.273]
τ_β	Between-analysis SD (slopes)	0.1021 [0.0995, 0.1047]
τ_{study}	Study heterogeneity SD	0.490 [0.479, 0.501]
$\text{ICC}_{\text{study}}$	Study ICC	0.775 [0.760, 0.790]

Notes: Panel (a) shows the estimated average treatment effect trajectory from the hierarchical model pooling all analyses with crossed study random effects. All effect sizes are expressed in standard deviation (SD) units; ratio measures (odds ratios, risk ratios) are converted using the probit transformation (Appendix E). The solid line indicates the posterior mean; the shaded region shows the 95% credible interval. The dashed horizontal line marks the null effect. Panel (b) reports the posterior mean and 95% credible interval for key model parameters: μ_α is the mean initial effect (SD), μ_β is the mean fadeout rate (SD per month), ρ is the correlation between initial effects and fadeout rates, τ_α and τ_β capture between-analysis heterogeneity, τ_{study} is study-level heterogeneity, and the ICC measures the fraction of variance at the study level.

3.3 Heterogeneity by Intervention Type

While the aggregate fadeout estimate is close to zero, this average may mask meaningful variation across intervention types. We test whether different categories of psychosocial intervention show different fadeout rates using inverse-variance weighted regression within each intervention cluster.[5]

For each of nine intervention clusters (excluding Breastfeeding Support due to insufficient sample size), we estimate:

$$Y_{it} = \alpha + \beta \cdot t_{it} + \varepsilon_{it} \tag{7}$$

[4]The crossed study random effects reveal that most heterogeneity in effect sizes operates at the study level: the study ICC is 0.78 (95% CI: $[0.76, 0.79]$), meaning that most of the non-slope variance is attributable to study-level factors (sample quality, implementation fidelity, context) that affect all outcomes measured within a study. This validates the importance of accounting for cross-analysis dependence when the same study contributes to multiple analyses.

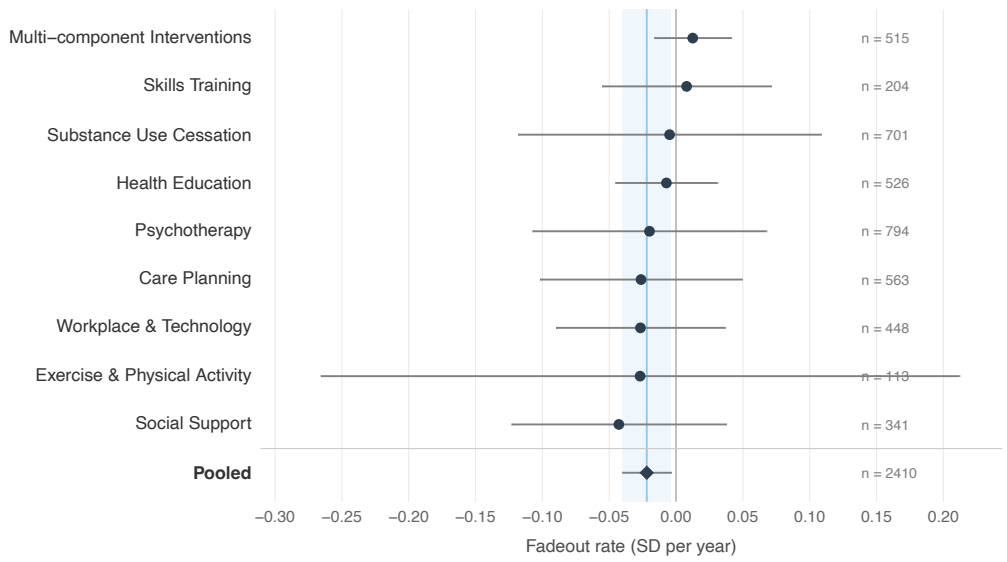
[5]Intervention clusters are defined by k-means clustering of intervention term embeddings extracted from review metadata (Population, Intervention, Comparison, Outcome descriptors), as described in Appendix C.

where Y_{it} is the effect size (in SD units) for observation i at time t_{it} (in years since treatment completion), with weights proportional to $1/\text{SE}_{it}^2$. Standard errors are clustered at the study level using CR2 small-sample correction. The sample is restricted to analysis groups meeting the same inclusion criteria as the main analysis (IQR of follow-up time ≥ 3 months and ≥ 3 studies).

Figure 3 presents the results. The pooled estimate across all clusters is $\hat{\beta} = -0.022$ SD/year (95% CI: $[-0.040, -0.004]$, $p = 0.025$)—larger in magnitude than the hierarchical model's estimate of 0.01 SD/year. The difference reflects the fact that this simple weighted regression does not apply the partial pooling and shrinkage of the Bayesian hierarchical model, and is estimated on a different (more restrictive) sample. Both estimates point to small fadeout. Critically, a test for heterogeneity across clusters fails to reject equality of slopes ($Q = 11.9$, $df = 8$, $p = 0.156$). No individual cluster shows statistically significant fadeout at the 5% level: point estimates range from −0.043 SD/year (Social Support) to +0.012 SD/year (Multi-component Interventions), but all confidence intervals include zero.

The uniformity across intervention types is notable. Fadeout rates do not differ significantly across psychotherapy, patient education, exercise programs, care management, and other intervention modalities—suggesting that the mechanisms underlying effect persistence operate similarly across diverse treatments.

Figure 3: *Fadeout Rates by Intervention Type*



Notes: Each row shows the estimated fadeout rate (change in effect size per year, in SD units) for one intervention cluster, estimated via inverse-variance weighted regression. The diamond shows the pooled estimate across all clusters; the blue vertical line and shaded band show its 95% confidence interval. Negative values indicate fadeout (declining effects over time). Standard errors clustered at the study level using CR2 small-sample correction with Satterthwaite degrees of freedom. Sample restricted to analysis groups with IQR of follow-up time ≥ 3 months and ≥ 3 studies, matching the main analysis sample. N indicates number of unique studies in each cluster.

4 Selection

A key threat to our persistence estimates is that follow-up timing is endogenous: researchers who observe persistence may invest in longer follow-up, while those who expect fadeout may stop measuring early. We address this concern through four complementary approaches, moving from direct tests of selection on observables to bounding and instrumenting for unobservable selection.

4.1 Selection on Short-run Effects

Researchers may choose to measure long-run outcomes only when they observe promising short-run results, biasing persistence estimates upward. We test for this selection using within-analysis comparisons between studies with and without long-run follow-up.

We classify studies within each analysis based on their temporal coverage relative to that analysis's median follow-up time. *Studies with long-run follow-up* have observations at both short-run ($\leq$ median) and long-run ($>$ median) time points; they provide the within-study variation that identifies fadeout. *Short-run-only studies* have observations only at short-run time points. If selection on early results drives long-run follow-up, studies with long-run data should have systematically higher short-run effects than short-run-only studies within the same analysis.

The regression sample includes only studies with at least one short-run observation. We estimate:

$$Y_{it}^{\text{short}} = \gamma \cdot \mathbf{1}[\text{has long-run}] + \alpha_j + \epsilon_{it} \tag{8}$$

where Y_{it}^{short} is the short-run effect size, α_j are analysis fixed effects, and standard errors are clustered by analysis. The coefficient γ measures whether studies with long-run follow-up have different short-run effects than short-run-only studies within the same analysis.

Figure 4 presents the results. Panel (a) shows the regression estimates. Without fixed effects, studies with long-run follow-up have marginally lower short-run effects (-0.028 SD, $p < 0.10$)—if anything, opposite to the selection prediction. With analysis fixed effects, the difference is smaller and statistically indistinguishable from zero (-0.009 SD, SE $= 0.015$). There is no evidence that studies with long-run follow-up had unusually promising short-run results.

Panel (b) provides a non-parametric visualization. We residualize short-run effects within each analysis and assign studies to deciles based on their residual effect size. The figure plots the probability of having long-run follow-up by decile. If selection drove long-run follow-up, this probability should increase with short-run effect size—studies with larger-than-expected effects should be more likely to have long-run measurements. Instead, the pattern is flat: studies in the highest decile (D10) are no more likely to have long-run follow-up than those in the lowest decile (D1). The overall rate of 18.1% is nearly constant across the distribution.

These findings suggest that selection into long-run follow-up is not strongly driven by short-run effect magnitudes. Researchers appear to choose follow-up timing based on factors other than early results—perhaps study design, funding, or outcome domain—rather than selectively continuing only

the most promising studies.

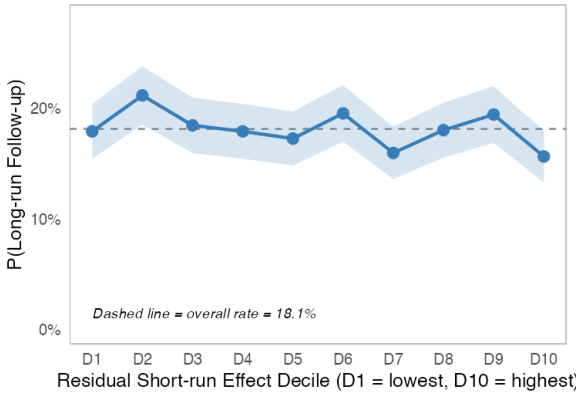
Figure 4: *Selection Test: Panel Studies vs. Short-run Effects*

(a) Regression Estimates

	(1) No FE	(2) Analysis FE
Has Long-run Follow-up	−0.028*	−0.009
	(0.015)	(0.015)
Num.Obs.	10 931	10 931
R2	0.001	0.346
Analysis FE	No	Yes
Pct. with long-run	19.6%	19.6%

* p < 0.1, ** p < 0.05, *** p < 0.01

(b) P(Long-run Follow-up) by Effect Decile



Notes: Sample restricted to studies with at least one short-run observation. Panel (a) regresses short-run effect sizes on an indicator for having long-run follow-up, with and without analysis fixed effects. Standard errors clustered by analysis. Panel (b) shows the probability of having long-run follow-up by decile of residual short-run effect (after partialling out analysis fixed effects). If selection on short-run results drove long-run follow-up, higher deciles should have higher probabilities; instead, the pattern is flat.

4.2 Correcting for Selection on Observables

The previous subsection found no evidence that studies with long-run follow-up had systematically different short-run effects. But selection could operate through other channels: perhaps analyses with certain characteristics (e.g., larger samples, more precise estimates, particular intervention types) are more likely to include long-run measurements, and these characteristics also correlate with persistence. We use inverse probability weighting (IPW) to correct for selection on this broader set of observables.

Selection into long-run follow-up can operate at two levels: some *analyses* never include any long-run measurements, while within analyses that do, some *studies* contribute long-run observations and others do not. We address both levels using separate IPW corrections that follow the logic of standard attrition weighting. A key constraint of IPW is that we can only reweight the observed population—analyses or studies that *do* have long-run follow-up—to better represent the full population.

Analysis-level correction. We first ask: are analyses with long-run follow-up systematically different from those without? Following standard attrition correction, we model the probability that an analysis has any observation beyond six months as a function of observable characteristics: mean short-run effect size, precision (mean SE), number of observations, effect measure type, and intervention cluster indicators (Appendix I). Analyses with long-run data that have characteristics *atypical* of long-run analyses—those unlikely to have been followed long-term based on observables—receive higher weight. Formally, the stabilized weight is $w_j = P(\text{long-run})/P(\text{long-run} \mid X_j)$, where the numerator is the marginal probability and the denominator is the propensity score. Figure I10 shows that propensity scores overlap substantially between analyses with and without long-run follow-up,

supporting the validity of IPW reweighting.

Table 2 presents results from both corrections. Panel A shows analysis-level results, aggregating slope estimates across the 1,184 analyses with long-run follow-up (analogous to the meta-meta-analysis in Section 3.2, but restricted to this subset). The unweighted mean slope is 0.0005 SD/month (SE = 0.0009), essentially zero. Precision weighting (by $1/SE^2$) yields −0.0002 SD/month. With IPW correction, the estimated slope becomes −0.0007 SD/month (95% CI: [−0.0013, −0.0000]). The IPW correction thus reveals slightly *more* fadeout than uncorrected estimates: analyses with characteristics typical of long-run follow-up tend to show less fadeout, so correcting for this selection increases the estimated average decline.

Study-level correction. We next ask: within analyses, are studies with long-run follow-up systematically different? Here, "long-run" means having observations beyond the analysis's median follow-up time. We model the probability that a study has long-run observations as a function of study characteristics and re-estimate the aggregate fadeout model with stabilized IPW weights (Figure I11 shows propensity overlap and weight distributions).

Panel B of Table 2 shows study-level results using a different approach: a pooled regression of effect sizes on follow-up time across all observations from studies with long-run data. Across 2,987 studies, the unweighted slope is −0.0033 SD/month. This is larger in magnitude than the hierarchical model estimate in Section 3.2 (0.0008 SD/month), reflecting that the pooled regression does not apply shrinkage toward zero as the hierarchical model does. Precision weighting yields −0.0025 SD/month, with IPW correction having minimal additional effect (−0.0025 SD/month). Unlike the analysis level, selection correction at the study level does not substantially change the estimated fadeout rate.

Table 2: *IPW Correction for Selection into Long-Run Follow-Up*

Panel A: Analysis-Level Correction			
Method	Mean Slope	SE	95% CI
Unweighted	0.0005	(0.0009)	[−0.0013, 0.0024]
Precision-weighted	−0.0002	(0.0002)	[−0.0006, 0.0003]
IPW + Precision	−0.0007	(0.0003)	[−0.0013, −0.0000]
N Analyses		1,184	

Panel B: Study-Level Correction			
Method	Mean Slope	SE	95% CI
Unweighted	−0.0033	(0.0017)	[−0.0066, −0.0000]
Precision-weighted	−0.0025	(0.0005)	[−0.0034, −0.0016]
IPW + Precision	−0.0025	(0.0005)	[−0.0034, −0.0016]
N Studies		2,987	

Notes: Panel A: sample restricted to 1,184 analyses with at least one observation beyond 6 months. Panel B: sample restricted to 2,987 studies with observations beyond the analysis median follow-up time. "Precision-weighted": weighting by $1/SE^2$. "IPW + Precision": combines precision weighting with inverse probability weighting using stabilized weights $w = P(\text{long-run})/P(\text{long-run} \mid X)$. Both panels use propensity models including short-run effect size, precision, sample size, and effect measure type. Weights trimmed at 95th percentile.

Interpretation. Selection correction does not reveal hidden fadeout. The corrected estimates (−0.0007 to −0.0025 SD/month) remain consistent with our main finding of limited average fadeout. If anything, analyses and studies most likely to include long-run measurements—based on observable characteristics—show *less* decline, so correcting for this selection slightly *increases* the estimated fadeout rate.

The limitation of IPW is that it cannot correct for selection on unobservables. If researchers choose follow-up timing based on private information about expected persistence—for instance, observing early signs of fadeout and deciding not to measure further—our estimates would remain biased. Nevertheless, the finding that observable selection operates in the *opposite* direction of the persistence hypothesis provides some reassurance: there is no evidence that selection inflates our persistence estimates.

4.3 Bounding Unobservable Selection

The previous subsection corrected for selection on observable characteristics, but a more fundamental concern is that researchers may endogenously choose follow-up periods based on their priors about effect persistence. If investigators expect an intervention's effects to fade quickly, they have an incentive to measure outcomes sooner—before the effects disappear. Conversely, if they expect effects to persist, they may invest in longer follow-up to document durable benefits. This creates a systematic bias: interventions expected to fade are measured at shorter horizons, so we never observe their fadeout; interventions expected to persist are measured at longer horizons, contributing disproportionately to our long-run estimates. The result is that our data over-represents persistence and under-represents fadeout, biasing estimates toward smaller decline rates. We cannot correct for this selection directly, but we can bound its potential impact by asking: how large could fadeout be under pessimistic assumptions about what we would have observed had researchers measured outcomes at longer horizons?

We construct pessimistic bounds by augmenting the data with synthetic observations that assume complete fadeout for studies lacking long-run measurements. For each study without observations beyond 24 months, we add a synthetic observation at 36 months with effect size equal to zero (the null effect). This assumes the worst case: that all missing long-run data would have shown complete effect disappearance. We then re-estimate our hierarchical model on the augmented data, yielding an upper bound on the fadeout rate. Details of the procedure, including how we assign standard errors to synthetic observations, are provided in Appendix J.

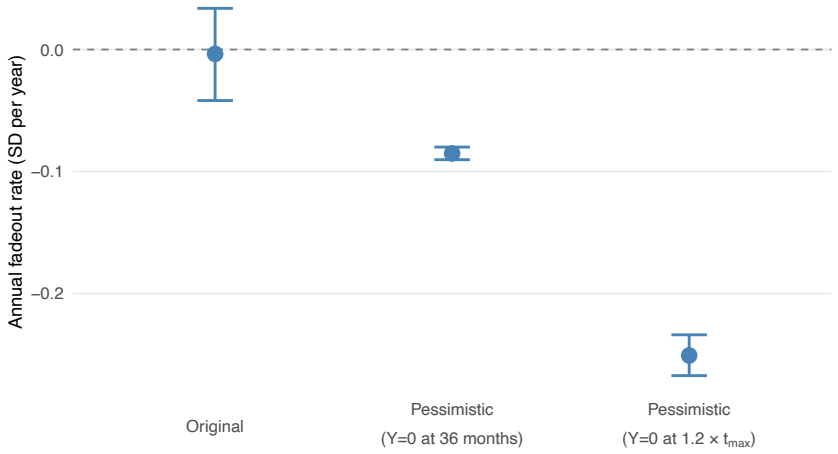
Figure 5 compares the estimated annual fadeout rates under our main specification and the pessimistic bounds. The original estimate shows an annual rate of approximately 0.01 SD per year, statistically indistinguishable from zero and slightly positive. Under the pessimistic bound, assuming complete fadeout for studies with missing long-run data, the estimated annual decline increases to 0.09 SD per year.

As an extreme sensitivity check, we also implement a more aggressive bound that adds synthetic null-effect observations for *all* studies (not just those missing long-run data) at 20% beyond each

analysis's maximum observed follow-up time. Under this specification, annual fadeout increases to 0.25 SD per year. This bound is deliberately extreme and likely overstates true fadeout, as it assumes every study would show complete effect disappearance shortly after its last observation.

These bounds bracket a meaningful range. Under the original specification, an intervention with an initial effect of 0.34 SD would retain most of its effect over several years. Under the pessimistic 36-month bound, approximately one-quarter of the initial effect would fade within the first year—not negligible, but still implying meaningful persistence. Under the extreme $1.2 \times t_{\max}$ bound, most of the initial effect would disappear within one year. We view the 36-month bound as the more informative of the two: it makes a strong but not implausible assumption about what missing data would have shown. The truth likely lies between the original estimate and this bound, implying annual fadeout somewhere in the range of 0.01–0.09 SD per year.

Figure 5: *Pessimistic Bounds on Fadeout*



Notes: Point estimates and 95% credible intervals for the annual fadeout rate under three specifications. "Original": main estimates from Section 3.2. The two pessimistic bounds add synthetic $Y = 0$ observations: at 36 months for studies without data beyond 24 months, or at $1.2 \times t_{\max}$ (20% beyond the maximum observed follow-up) for all studies. The dashed line marks zero (no fadeout).

4.4 Instrumental Variable: Funding Shocks

The previous subsections addressed selection using observables (IPW) and worst-case assumptions (bounds). We now use an exogenous instrument for follow-up duration to estimate the magnitude and direction of selection bias in the slope. If follow-up timing is endogenous (driven by researchers' expectations about persistence), an instrument that shifts follow-up length for reasons unrelated to treatment effect dynamics can identify the resulting bias and, via a control function approach, adjust for it.

Instrument. We construct an instrument from NIH disease-area funding. The National Institutes of Health tracks annual spending across disease and condition categories through its Research, Condition, and Disease Categorization (RCDC) system. We match each Cochrane review to RCDC categories using the cosine similarity between text embeddings of the review abstract and category

descriptions, converted to weights via softmax (Appendix O). The instrument for study i ending in year t is

$$IV_{it} = \sum_k w_{ik} \log(\text{funding}_{k,t}) \tag{9}$$

where w_{ik} is the match weight between Cochrane review i and RCDC category k, and $\text{funding}_{k,t}$ is NIH spending on category k in year t. The instrument is standardized to a z-score (demeaned and divided by its standard deviation), so coefficients are interpretable as the effect of a one-SD increase in matched-area log funding. The identifying variation comes from year-to-year fluctuations in NIH categorical spending (driven by budget events such as the American Recovery and Reinvestment Act (2009) and sequestration (2013)), interacted with the disease-area composition of each Cochrane review.

Identification. The exclusion restriction requires that funding shocks affect follow-up duration but not the post-treatment decay rate of treatment effects. Our measure of follow-up, t, is time from the *end of treatment*, not from randomization. Funding that improves study implementation (through better-resourced data collection, larger samples, or stronger treatment fidelity) affects the end-of-treatment effect size θ_i (the intercept) but not the rate at which effects decay after treatment concludes. Review fixed effects absorb persistent differences across disease areas that might correlate with both funding levels and effect persistence. For the exclusion restriction to fail, funding shocks would need to differentially affect the *post-treatment decay rate* of studies within the same Cochrane review, a channel we regard as implausible.

First stage. The instrument predicts follow-up duration. Table 3 reports first-stage regressions on 1,704 studies with known end dates in the 2010–2022 window. Within a Cochrane review and year, a one-SD increase in matched-area funding in the year the study ends raises the probability of follow-up exceeding one year by 18.1 percentage points, off a base rate of 13.38%. The effect strengthens to 29.3 percentage points when we add review-specific linear time trends. To test which timing of funding matters, the table enters five single-year funding windows ($t-2$ through $t+2$) jointly: only contemporaneous funding (year t) is consistently significant, while pre-treatment and post-treatment windows are individually insignificant. This pattern is consistent with funding operating through study planning and implementation rather than through confounding trends. Figure O27 in Appendix O confirms non-parametrically that studies in higher funding terciles have longer follow-up, and Table O28 tests for differential effects by country.

4.4.1 Selection Correction

Researchers who expect persistence have reason to follow up longer; if so, the studies contributing long-run data disproportionately represent persistent effects, and the naive slope understates true fadeout. The bias arises from a compositional shift: at short horizons, all studies contribute observations, but at long horizons only studies with long follow-up remain. If those surviving studies have systematically flatter slopes, the estimated average fadeout rate is biased toward zero. We use

Table 3: *First Stage: Funding Instrument and Follow-Up Duration*

	P(≥90d)		P(≥180d)		P(≥270d)		P(≥365d)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Funding $t-2$	0.022	0.019	0.030	0.015	0.024	0.028	0.008	-0.007
	(0.026)	(0.039)	(0.028)	(0.040)	(0.025)	(0.037)	(0.022)	(0.035)
Funding $t-1$	0.093***	0.046	0.065	0.037	0.076**	0.083*	0.004	0.005
	(0.036)	(0.048)	(0.041)	(0.053)	(0.037)	(0.049)	(0.032)	(0.045)
Funding t	0.075	0.004	0.064	0.133	0.146**	0.286**	0.181***	0.293**
	(0.079)	(0.133)	(0.077)	(0.137)	(0.069)	(0.144)	(0.067)	(0.140)
Funding $t+1$	-0.075	-0.133	-0.052	-0.162	-0.095	-0.098	-0.091	-0.181*
	(0.084)	(0.128)	(0.083)	(0.130)	(0.070)	(0.115)	(0.063)	(0.105)
Funding $t+2$	0.053	-0.136	0.049	0.034	-0.000	0.064	0.021	0.028
	(0.081)	(0.126)	(0.077)	(0.128)	(0.063)	(0.121)	(0.057)	(0.113)
N	1,704	1,704	1,704	1,704	1,704	1,704	1,704	1,704
R^2	0.418	0.502	0.439	0.518	0.318	0.401	0.327	0.405
Mean dep. var.	0.483	0.483	0.356	0.356	0.178	0.178	0.134	0.134
Review + Year FEs	✓	✓	✓	✓	✓	✓	✓	✓
Review × trend		✓		✓		✓		✓

Notes: The unit of observation is a study (unique trial within a Cochrane review). The dependent variable is an indicator for whether the study's maximum post-treatment follow-up exceeds the stated threshold (90, 180, 270, or 365 days from end of treatment). Each column reports a linear probability model with five single-year funding instruments entered jointly. The instrument is $IV_{it} = \sum_k w_{ik} \log(\text{funding}_{k,t+s})$, where w_{ik} are softmax weights from embedding similarity between Cochrane review i and NIH RCDC category k, and $s \in \{-2, -1, 0, +1, +2\}$ indexes the funding year relative to study end year t. Odd columns include review and year fixed effects; even columns additionally include review-specific linear time trends. HC1 standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

a control function to adjust for this selection and recover the corrected fadeout rate.

First stage. The idea is to use the funding instrument to predict which studies will have long follow-up, then control for this predicted selection in the fadeout regression. We model the probability that a study has any follow-up beyond one year as a linear function of the funding instrument:

$$F_i = \beta \cdot IV_{it} + \gamma_r + \delta_t + \eta_i \tag{10}$$

where $F_i = \mathbf{1}[\text{max follow-up} \geq 365\text{d}]$, γ_r are review fixed effects, δ_t are year fixed effects, and IV_{it} is the contemporaneous funding instrument from (9). This is a linear probability model; its residuals $\hat{\eta}_i$ capture each study's deviation from the follow-up duration predicted by its funding environment. A study with $\hat{\eta}_i > 0$ has longer follow-up than its funding would predict, suggesting the researcher chose to follow up for reasons other than funding, plausibly because they expected persistence.

Second stage. We add the first-stage residuals and their interaction with follow-up time to the fadeout regression:

$$Y_{it}^{\text{flip}} = \alpha_1 t + \delta_\mu \hat{\eta}_i + \delta_\nu \hat{\eta}_i \cdot t + \gamma_j + \varepsilon_{it} \tag{11}$$

where Y_{it}^{flip} is the direction-standardized effect size, t is years from the end of treatment, and γ_j are analysis fixed effects. The two control function terms serve distinct purposes. The level term $\hat{\eta}_i$ absorbs any difference in average effect sizes between studies with endogenously long versus short

follow-up (for instance, if researchers follow up on interventions with larger initial effects). The interaction $\hat{\eta}_i \cdot t$ absorbs the more consequential bias: differences in *fadeout rates* between these groups. If $\delta_\nu > 0$, studies that chose long follow-up (conditional on funding) show less fadeout than those that did not, confirming that endogenous follow-up timing biases the naive slope toward zero. After conditioning on both terms, α_1 is the selection-corrected fadeout rate: what we would estimate if follow-up timing were determined entirely by funding shocks rather than by researchers' expectations about persistence.

As a robustness check, we replace the binary first stage with a discrete-time hazard model that divides follow-up into 90-day bins and estimates a bin-level LPM for the probability of exiting follow-up, allowing the effect of funding to vary by bin. The residuals $\hat{\eta}_{ik}$ from this model vary across observations within a study, providing a horizon-specific selection correction rather than the single study-level value $\hat{\eta}_i$ (Appendix O).

Table 4: *Selection-Corrected Fadeout Rate*

	(1) Naive	(2) Linear CF	(3) Discrete-Time Hazard CF
Follow-up time (years)	-0.069***	-0.099**	-0.099***
	(0.026)	(0.043)	(0.035)
CF term (level)		-0.095*	-0.111
		(0.053)	(0.068)
CF term × time		0.134*	0.182*
		(0.070)	(0.104)
N	8,961	8,961	8,944
R^2	0.605	0.606	0.607
Analysis FE	✓	✓	✓
Bootstrap SE (study)	✓	✓	✓

Notes: The unit of observation is a study–time point. The dependent variable is the direction-standardized effect size Y_{it}^{flip} (SD units). Column (1) estimates the naive fadeout rate. Column (2) adds a linear control function: LPM residuals $\hat{\eta}_i$ from a first-stage regression of $\mathbf{1}[\text{max follow-up} \geq 365\text{d}]$ on the contemporaneous funding instrument, with review and year fixed effects (first-stage coefficient: 0.125, SE = 0.059). Column (3) replaces the study-level CF with bin-level LPM residuals $\hat{\eta}_{ik}$ from a discrete-time hazard model (90-day bins), matched to each observation's follow-up horizon. The "CF term × time" coefficient ($\hat{\delta}_\nu$) tests whether selection biases the slope; $\hat{\delta}_\nu > 0$ indicates that studies selected for long follow-up show less fadeout. Bootstrap standard errors (1,000 iterations, resampling studies) in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Results. Table 4 reports the results, estimated on 1,472 studies (8,961 observations) in the instrument subsample. Standard errors are bootstrapped (1,000 iterations, resampling studies with replacement and re-estimating both stages) to account for estimation uncertainty in the first-stage residuals. Column (1) shows the naive fadeout rate: −0.069 SD per year (SE = 0.026), consistent with the full-sample estimate from Section 3. Column (2) adds the linear control function. The corrected slope steepens to −0.099 SD per year, and the selection test coefficient is $\hat{\delta}_\nu = 0.134$ (SE = 0.070, $p = 0.055$), positive as predicted, indicating that studies selected for long follow-up show less fadeout. Column (3) uses the discrete-time hazard control function, yielding a corrected slope of −0.099 SD per year with $\hat{\delta}_\nu = 0.182$ ($p = 0.073$), the same sign but less precisely estimated.

The correction amounts to 0.030 SD per year: the naive estimate understates fadeout by roughly

two-fifths. Selection does appear to bias the slope toward zero, consistent with researchers investing in longer follow-up when they expect durable effects. However, the corrected fadeout rate of approximately 0.1 SD per year remains modest relative to initial effects of 0.34 SD. At this rate, a typical intervention would retain roughly 70% of its initial effect after one year. Selection explains only a fraction of the apparent persistence: even after correction, psychosocial intervention effects appear durable.

The control function subsample covers approximately 17% of the full estimation sample, restricted to studies with known calendar end dates in the 2010–2022 RCDC funding window. Generalizability to the full sample depends on the assumption that selection operates similarly across funding contexts.

5 Robustness

We test three alternative explanations for our persistence finding: low statistical power, non-linear functional form, and within-study temporal correlation.

5.1 Statistical Power

The low rate of detectable dynamics (5.5%) could reflect insufficient statistical power rather than genuine persistence. If power explains our results, analyses with better statistical properties should detect fadeout more frequently. We test this prediction using simulation.

For each of 163 high-quality analyses—those with at least ten studies and an interquartile range of follow-up times of at least three months—we simulate statistical power under a range of true fadeout rates (Appendix H). Each simulation uses the analysis's empirical configuration: its number of studies, proportion of panel studies, follow-up time range, and standard error distribution. We generate data with known true slopes (ranging from 0.01 to 0.08 SD per month), fit our Bayesian hierarchical model, and record whether the 90% credible interval excludes zero. The minimum detectable effect (MDE) is the smallest slope at which simulations achieve 80% power.

Figure 6 presents the results. Panel (a) shows simulated power: at moderate slopes of 0.03 SD per month, 60–80% of analyses achieve 80% power depending on heterogeneity assumptions.[6] At larger slopes of 0.05 SD per month, this rises to 80–90%. Panel (b) compares simulated power with empirical detection rates. If power explained our null results, we would expect higher detection rates among better-powered analyses. Instead, the pattern is flat or inverted: analyses with the best power (MDE = 0.02) detect dynamics 12% of the time, comparable to the 14% among the worst-powered analyses (MDE > 0.08).

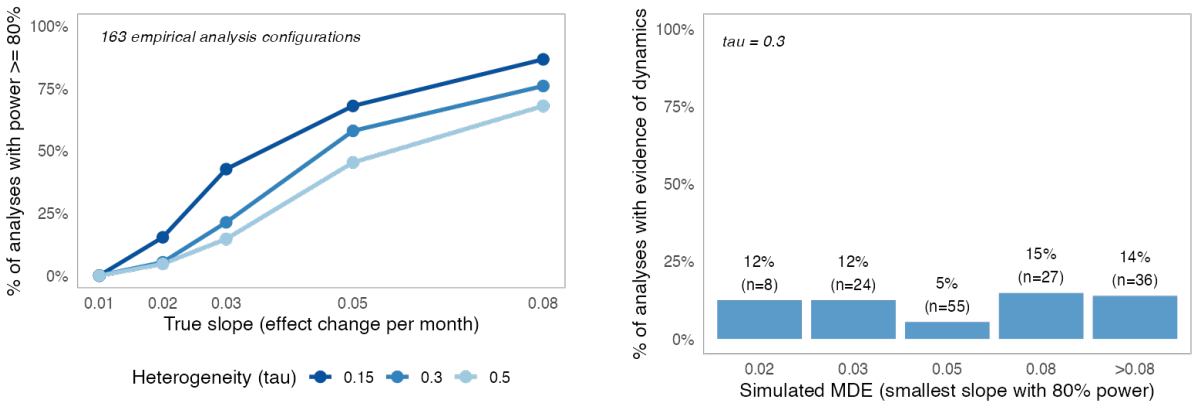
This flat pattern is inconsistent with the power explanation. If fadeout effects were common but simply undetected due to low power, better-powered analyses should find more fadeout—yet they do not. The most parsimonious interpretation is that true fadeout effects are genuinely small, so even well-powered analyses rarely detect them. The simulation also provides a useful benchmark:

[6] We simulate power under three between-study heterogeneity values: $\tau \in \{0.15, 0.30, 0.50\}$. The median τ estimated across our meta-analyses is 0.27, with an interquartile range of 0.19–0.41. Our simulation values thus span the observed distribution from below the 25th percentile to above the 75th percentile.

our high-quality analyses would have detected fadeout at a rate of 0.03 SD/month with 60–80% probability, suggesting we can rule out fadeout of this magnitude or larger with reasonable confidence.

Figure 6: *Statistical Power Analysis*

(a) Simulated Power by True Slope **(b) Empirical Detection vs Simulated Power**



Notes: Panel (a) shows the proportion of 163 high-quality analyses achieving 80% power at each true slope value, across three assumptions about between-study heterogeneity ($\tau \in \{0.15, 0.30, 0.50\}$, chosen to span the empirical distribution where median $\tau = 0.27$). Panel (b) shows the empirical detection rate (proportion with evidence of dynamics) for analyses grouped by their simulated minimum detectable effect (MDE)—the smallest slope at which simulations achieved 80% power, assuming $\tau = 0.3$. If low power explained our null results, detection rates should increase as MDE decreases (moving left); instead, the pattern is flat or inverted.

5.2 Robustness to Functional Form

Our main specification assumes linear fadeout, but this may mask non-linear patterns such as rapid early decay followed by stabilization, or delayed effects that strengthen before fading. We test robustness using two alternative specifications: a piecewise linear spline (with knots at 3, 6, and 12 months) that allows the fadeout rate to vary across time segments, and an exponential decay model $Y(t) = \alpha e^{-\lambda t} + c$ that permits asymptotic persistence at a non-zero level. Both models are estimated at the individual analysis level (1,503 analyses) and in the aggregate meta-meta-analysis.

All three specifications yield similar trajectories: an initial effect of approximately 0.34 SD that remains largely stable over time. At the meta-meta level, 12-month predictions range from 0.33–0.36 SD across models. The spline model shows no statistically significant slope changes at any knot, and analysis-level slopes are highly correlated across specifications. These results confirm that the linear specification is adequate; full details, parameter estimates, and comparison figures are reported in Appendix K.

5.3 Robustness to Within-Study Correlation

Our baseline model estimates a common slope β_j across studies within each analysis, treating multiple time points from the same panel study as independent observations conditional on the study intercept. If within-study residuals are positively correlated, this overstates the effective information content of panel observations and may understate uncertainty about the slope. We re-estimate the model with study-specific random slopes $\beta_{ij} \sim N(\beta_j, \tau^2_\beta)$, so that each panel study contributes a

single noisy slope estimate rather than multiple independent observations. Across 1,484 matched analyses, population slope estimates are highly correlated between the two models ($r = 0.77$), with nearly identical means. The random slopes model produces wider credible intervals—1.6 times the baseline width at the median—reducing the number of analyses with detectable dynamics from 80 to 14, but revealing no additional fadeout. Full details are reported in Appendix L.

5.4 Publication Bias

If selective reporting is more severe for short-run than long-run outcomes, initial effect sizes would be inflated, producing apparent fadeout that partly reflects publication bias rather than genuine dynamics. Standard diagnostics detect some bias: caliper tests show modest p-value bunching at conventional thresholds, and Egger's test reveals funnel asymmetry within analyses. A PET-PEESE specification (Stanley & Doucouliagos, 2014) with an SE × time interaction confirms that the bias is time-varying—less precise observations show steeper apparent fadeout ($p = 0.037$), consistent with significance filtering that inflates short-run effects in underpowered studies. However, correcting for this bias barely changes the estimated fadeout slope: the PEESE-corrected estimate is virtually identical to the naive, and the interaction specification, if anything, reduces estimated fadeout. Full results are reported in Appendix N.

6 Discussion

This paper asks whether psychosocial interventions fade out. Drawing on 1,503 analyses from Cochrane systematic reviews, we find that fadeout is small. Only 5.5% of analyses show detectable time trends, and the average fadeout rate from our hierarchical model is approximately 0.01 SD per year, pooling across 4,738 analyses. This is economically modest relative to mean initial effects of 0.34 SD. Pessimistic bounds that assume complete fadeout for studies lacking long-run data yield an upper bound of approximately 0.09 SD per year. An instrumental variable approach using NIH funding shocks suggests that selection into long-run follow-up understates fadeout by roughly 0.030 SD per year, yielding a corrected estimate of approximately −0.099 SD per year—still modest relative to initial effects. These findings are robust to concerns about statistical power, selection on observables, alternative functional forms, and within-study correlation.

These results have direct implications for how policymakers value social programs. Cost-benefit analyses of early childhood programs are acutely sensitive to persistence assumptions—the net present value of landmark programs like the Perry Preschool Project and Carolina Abecedarian depends critically on whether treatment effects are projected to persist, fade, or grow over the life cycle (García, Heckman, Leaf, & Prados, 2020; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010). The same sensitivity applies broadly. Under a pessimistic assumption of rapid fadeout (0.05 SD per month, as estimated for IQ-raising interventions by Protzko 2015), a typical psychosocial intervention's benefits would be exhausted within seven months. Under our estimate of approximately 0.01 SD per year, the model projects that most of the initial effect would persist beyond the first year—though this projection relies on a linear extrapolation from data concentrated at shorter horizons.

Even under our pessimistic bound (0.09 SD/year), roughly three-quarters of the initial effect persists after one year. These differences matter: at a 3% discount rate, shifting from rapid to gradual fadeout can increase the present value of benefits by a factor of three or more. For policymakers scaling psychosocial interventions—such as those recommended under the WHO's mhGAP guidelines for mental health in low- and middle-income countries—our estimates provide an empirical basis for the persistence assumptions that underpin cost-effectiveness calculations.

The finding that larger initial effects fade faster ($\rho = -0.35$) has practical implications for program evaluation. Interventions producing dramatic short-run improvements may require maintenance protocols or booster sessions to sustain gains. This pattern is consistent with Cunha and Heckman's (2007) technology of skill formation, in which the productivity of early investments depends on complementary later investments—without reinforcement, initial gains depreciate. It is also consistent with the broader finding that treatment effects are heterogeneous across contexts (Vivalt, 2020); our contribution is to show that this heterogeneity has a systematic temporal structure, with $\rho < 0$ implying mean reversion in effect sizes over time. Policymakers should be cautious about extrapolating short-run "miracle" results to long-run cost-effectiveness.

Several limitations warrant caution. First, follow-up data are concentrated at short horizons: the median follow-up is approximately 6.1 months, and only 1,047 observations (across 572 analyses) exceed two years, so our results are most directly supported within the first one to two years. Second, we cannot rule out selection on unobservables: researchers may discontinue measurement when effects fade, and our pessimistic bounds and instrumental variable correction address but do not eliminate this concern, placing the true fadeout rate in the range of 0.01–—0.099 SD per year (control function corrected) to 0.09 SD per year (pessimistic bound). Third, follow-up periods were extracted using large language models, which may introduce measurement error. Fourth, our sample draws on Cochrane reviews, which are predominantly clinical and health-focused; generalizability to education, labor market, and poverty-reduction interventions remains uncertain. Fifth, publication bias in the underlying studies could cut in either direction: selective publication of larger effects would overstate initial impacts and, given $\rho < 0$, inflate fadeout, while selective publication of persistent effects would overstate persistence (I. Andrews & Kasy, 2019; Kicinski, Springate, & Kontopantelis, 2015); Cochrane's search and inclusion protocols mitigate but do not eliminate this concern. Finally, we estimate average fadeout across a heterogeneous set of interventions; identifying which contexts produce more or less persistent effects remains an important direction for future work.

References

Almond, D., Currie, J., & Duque, V. (2018). Childhood circumstances and adult outcomes: Act II. *Journal of Economic Literature*, *56*(4), 1360–1446.

Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C. L., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders*, *55*, 70–78.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–2794.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fadeout of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, *21*(2), 55–97.

Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2016). Worms at work: Long-run impacts of a child health investment. *Quarterly Journal of Economics*, *131*(4), 1637–1680.

Barlow, J., Bergman, H., Kornør, H., Wei, Y., & Bennett, C. (2016). Group-based parent training programmes for improving emotional and behavioural adjustment in young children. *Cochrane Database of Systematic Reviews*(8).

Bhat, B., de Quidt, J., Haushofer, J., Patel, V., Rao, G., Schilbach, F., & Vautrey, P.-L. (2022). *The long-run effects of psychotherapy on depression, beliefs, and economic outcomes* (Working Paper No. 30011). National Bureau of Economic Research.

Blattman, C., Chaskel, S., Jamison, J. C., & Sheridan, M. (2023). Cognitive behavioral therapy reduces crime and violence over ten years: Experimental evidence. *American Economic Review: Insights*, *5*(4), 527–545.

Bouguen, A., Huang, Y., Kremer, M., & Miguel, E. (2019). Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*, *11*, 523–561.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions* (Working Paper No. 18038). National Bureau of Economic Research.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131.

Conti, G., Heckman, J. J., & Pinto, R. (2016). The effects of two influential early childhood interventions on health and healthy behaviour. *The Economic Journal*, *126*(596), F28–F65.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with

other treatments. *The Canadian Journal of Psychiatry*, *58*(7), 376–385.

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

García, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2020). Quantifying the life-cycle benefits of an influential early-childhood program. *Journal of Political Economy*, *128*(7), 2502–2541.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, *37*(1), 3–30.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., ... Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, *344*(6187), 998–1001.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage Publications.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*(1–2), 114–128.

Ioannidis, J. P. (2009). Integration of evidence from multiple meta-analyses: A primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ*, *181*(8), 488–493.

Kicinski, M., Springate, D. A., & Kontopantelis, E. (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine*, *34*(20), 2781–2793.

Kuyken, W., Warren, F. C., Taylor, R. S., Whalley, B., Crane, C., Bondolfi, G., ... others (2016). Efficacy of mindfulness-based cognitive therapy in prevention of depressive relapse: An individual patient data meta-analysis from randomized trials. *JAMA Psychiatry*, *73*(6), 565–574.

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). Large language models: An applied econometric framework. *arXiv preprint arXiv:2412.07031*.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, *11*(1), 57–91.

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*, *48*(7), 1815–1820.

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence*, *53*, 202–210.

Rith-Najarian, L. R., Mesri, B., Lane, M., Engstrom, M., & Chavira, D. A. (2019). Durability of cognitive behavioral therapy effects for youth and adolescents with anxiety, depression, or traumatic stress: A meta-analysis on long-term follow-ups. *Behavior Therapy*, *50*(1), 225–240.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Calculating the standardized mean difference for a contrast of means with unequal sample sizes, missing data, and/or variances. *Psychological Methods*, *8*(1), 94–106.

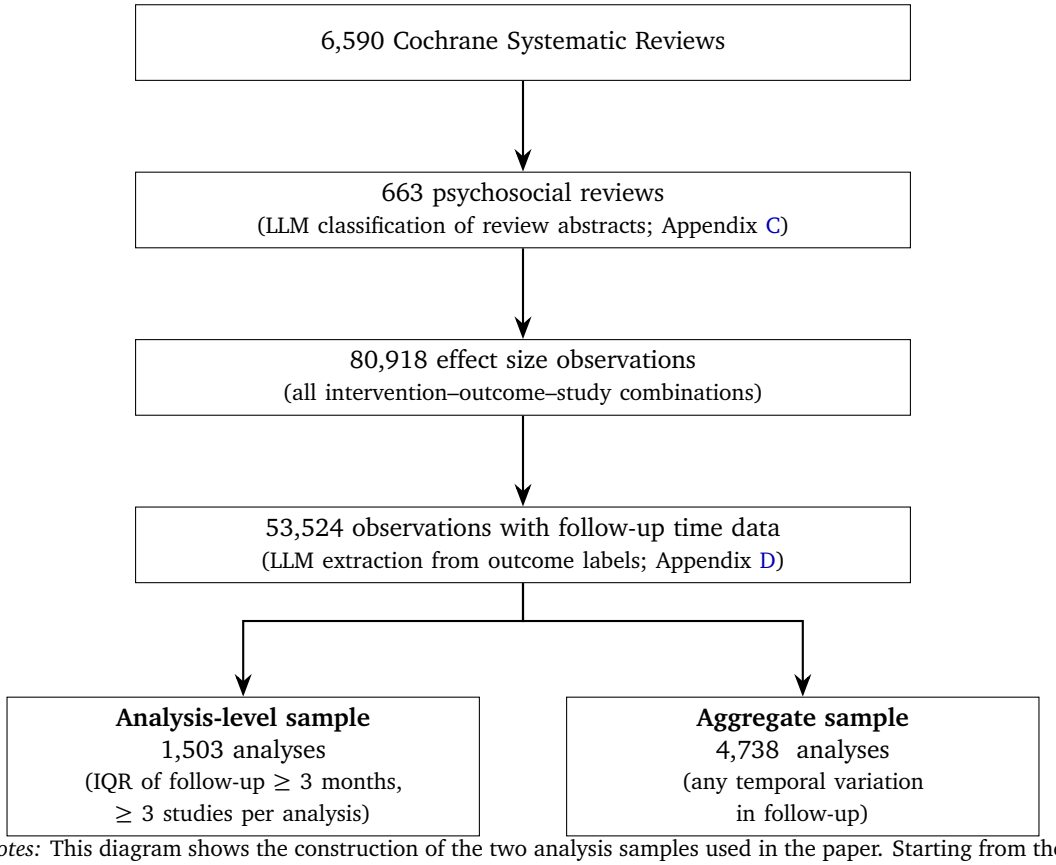
Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, *18*(6), 3045–3089.

APPENDIX

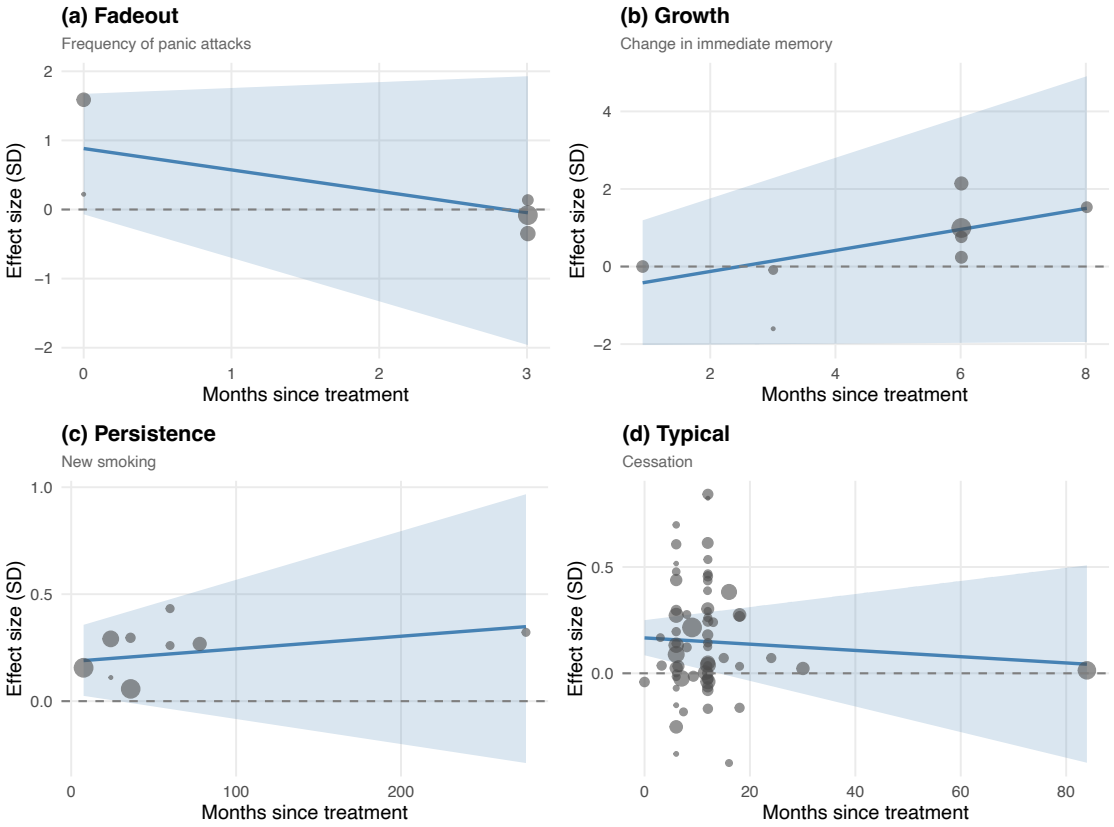
A Sample Construction

Figure A1: *Sample Construction Flow Diagram*



Notes: This diagram shows the construction of the two analysis samples used in the paper. Starting from the full Cochrane Database of Systematic Reviews, we identify psychosocial reviews using LLM classification, extract effect sizes and follow-up periods, and define analysis groups (unique intervention–outcome–comparison combinations). The analysis-level sample imposes stricter inclusion criteria to enable reliable estimation of individual-analysis fadeout rates; the aggregate sample uses all available temporal variation to maximize power for estimating the average fadeout rate.

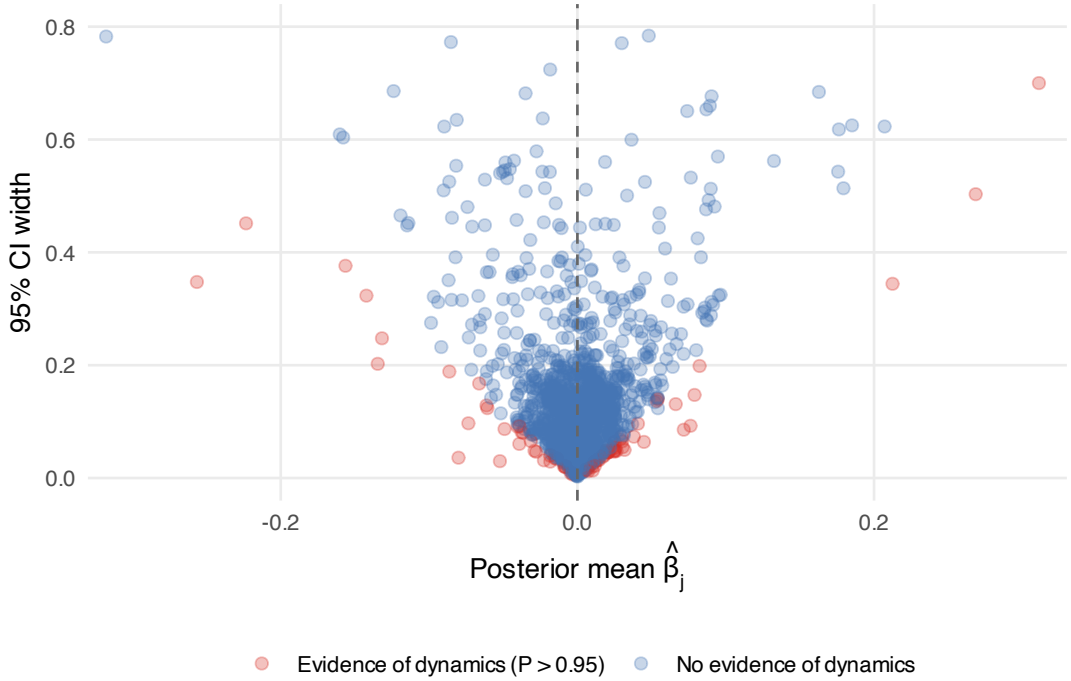
Figure A2: *Example Individual Analyses*



Notes: Each panel shows data from one individual analysis, illustrating the range of patterns in our sample. Points represent individual study effect sizes (in SD units), with point size proportional to precision ($1/SE^2$). The blue line shows the fitted linear trend from the Bayesian hierarchical model; the shaded band shows the 95% credible interval. Panel (a) shows an analysis with clear fadeout (declining effects over time). Panel (b) shows effect growth. Panel (c) shows a well-estimated near-zero slope (persistence). Panel (d) shows a typical analysis near the median slope.

B Supplementary Tables and Figures

Figure B3: *Slope Estimate vs Credible Interval Width*



Notes: Each point represents one analysis. Red points indicate analyses with evidence of dynamics (posterior probability of direction > 0.95). Analyses detecting dynamics tend to have larger absolute slope estimates and narrower credible intervals, consistent with the statistical requirement that effects must be precisely estimated to achieve high posterior probability of direction.

Table B4: *Outcome Types*

Category	Labels	Obs.	Representative Outcomes
Functional Outcomes & Cognition	1,210	16,205	Pain, Cessation, Disability
Anxiety & Depression Severity	929	15,964	Depression, Anxiety, Depressive symptoms
Adverse Events & Attrition	1,185	12,660	Remission, treatment acceptability (dropouts), Low & unclear attrition
Physical Health & Body Weight	908	9,160	Fatigue, Change in BMI z score, Body mass index (BMI) change
Substance Use & Health Behaviors	713	7,444	Abstinence, Smoking cessation, cessation
Parenting & Child Development	879	5,923	Child conduct problems: parent report, Psychosocial functioning, Child conduct problems: independent report
Maternal & Perinatal Health	532	3,899	Neonatal mortality, Spontaneous vaginal birth, Caesarean birth
Psychiatric Symptoms	613	3,705	Change in a global measure of cognition (composite), Change in a global measure of cognition, Mental state: General - clinically important change (no improvement)
Quality of Life	489	3,685	Quality of life, Health-related quality of life, Quality of Life
Healthcare Utilization	490	3,452	Hospital admissions, School Absences (mean days), Exacerbations leading to emergency department (ED) visits

Notes: Unique outcome labels (outclab_clean) clustered using k-means on text embeddings. "Labels" is the number of unique outcome labels in each cluster; "Obs." is the total number of effect-size observations across those labels. Representative outcomes are the three most common labels (by observation count) in each cluster. Labels with only temporal information (e.g., "6 months") are excluded.

C Psychosocial Intervention Selection

This appendix describes the procedure for classifying Cochrane reviews as psychosocial interventions. From an initial corpus of 6,590 Cochrane systematic reviews, we identified 663 reviews studying psychosocial interventions using a large language model (LLM) classification approach.

C.1 Classification Approach

We define *psychosocial interventions* as treatments that operate primarily through psychological, behavioral, or social mechanisms rather than pharmacological, surgical, or purely physical ones. This includes psychotherapy (cognitive behavioral therapy, dialectical behavior therapy), counseling and support interventions, psychoeducation, behavioral activation, mindfulness-based interventions, social skills training, parenting programs, and occupational therapy with a psychosocial focus. It excludes pharmacological treatments, surgical interventions, physical therapy aimed solely at physical rehabilitation, and diagnostic or screening procedures.

Classification was performed using GPT-4o (OpenAI) with temperature set to zero for reproducibility. For each Cochrane review, the model received the review title, abstract (truncated to 2,000 characters), keywords, and PICO (Population, Intervention, Comparison, Outcome) descriptors extracted from the Cochrane database. The model returned a structured JSON response containing a classification decision and one-sentence justification.

C.2 Classification Criteria

The LLM was instructed to classify each review into one of three categories:

- **YES**: The intervention is *clearly* psychosocial—it primarily targets mental, emotional, behavioral, or social well-being through psychological or social means.
- **NO**: The intervention is *clearly not* psychosocial—for example, medication, surgery, or physical therapy aimed solely at physical outcomes.
- **UNSURE**: The intervention is ambiguous, has mixed components, or its primary mechanism is unclear.

Table C5 provides examples from the classification prompt illustrating each category. Reviews classified as YES were included in the analytic sample; reviews classified as NO or UNSURE were excluded. We adopted a conservative approach, preferring false negatives (excluding borderline psychosocial interventions) to false positives (including non-psychosocial interventions).

C.3 Classification Prompt

The complete classification prompt is reproduced below:

Table C5: *Psychosocial Classification Examples*

Category	Example Interventions
YES (Psychosocial)	Cognitive behavioral therapy (CBT), dialectical behavior therapy (DBT), counseling, psychoeducation, social skills training, support groups, mindfulness-based interventions (for stress/emotion regulation), arts therapies, occupational therapy with psychosocial focus, vocational rehabilitation, activity scheduling, relaxation training, self-management programs, parenting training
NO (Not Psychosocial)	Medication, surgical procedures, post-surgical rehabilitation, physical therapy aimed solely at physical outcomes, dietary interventions without behavioral focus, screening and diagnostic procedures
UNSURE	Interventions with mixed pharmacological and psychological components, physical exercise programs (may target mental or physical health), complex multi-component interventions where the psychosocial element is secondary

Prompt: Classify the following Cochrane review:

- **"YES"**: The intervention is clearly psychosocial (primarily targets mental, emotional, behavioral, or social well-being through psychological/social means).
- **"NO"**: The intervention is clearly not psychosocial (e.g., medication, surgery, purely physical therapy).
- **"UNSURE"**: The intervention is ambiguous, has mixed components, or its primary aim is unclear.

Psychosocial (YES) Examples: Behavioral therapy (CBT, DBT), counseling, psychoeducation, social skills training, support groups, mindfulness (for stress/emotion), arts therapies, occupational therapy (psychosocial focus), vocational interventions, activity scheduling, relaxation, self-management, parenting training.

NOT Psychosocial (NO) Examples: Medication, ANY relation to surgery or post-surgery, physical therapy (solely for physical outcomes), dietary changes (unless primarily behavioral/emotional), screening/diagnosis.

[Review metadata inserted here: Title, Abstract, Keywords, Intervention, Population, Outcome]

Return your answer as valid JSON: `{"justification": "...", "relevant": "YES" | "NO" | "UNSURE"}`

C.4 Sample Characteristics

The classification procedure identified 663 Cochrane reviews as studying psychosocial interventions, representing approximately 8% of the initial corpus. These reviews span diverse intervention domains including mental health treatment, patient education, behavioral health, caregiver support, and rehabilitation. The distribution of intervention types within this sample is characterized further using the PICO clustering procedure described in Table 1.

D Follow-up Period Extraction

This appendix describes the procedure for extracting follow-up periods from Cochrane review data. A critical piece of information for our analysis is the *follow-up period*: the time between when treatment ends and when outcomes are measured. This information is rarely coded systematically in published meta-analyses, despite being essential for understanding whether intervention effects persist, fade, or strengthen over time.

D.1 Definition: Time from End of Treatment

We define the follow-up period as the number of days from the **end of treatment** to the outcome measurement—not from randomization or baseline. This distinction matters because we want to understand how effects persist *after* the intervention stops, not how long the study has been running.

For example, if a study randomizes participants, delivers an 8-week intervention, and measures outcomes at 6 months post-randomization, the follow-up period is approximately $183 - 56 = 127$ days (6 months minus 8 weeks). For studies with continuous or ongoing interventions where treatment does not end before outcome measurement, we code the follow-up period as zero.

D.2 Data Sources

Two data sources were used for extraction:

1. **Study characteristics:** Structured information extracted from the "Characteristics of included studies" tables in each Cochrane review, including descriptions of methods, participants, interventions (with timing and duration details), and outcomes (with assessment timepoints).
2. **Measurement labels:** The outcome label associated with each effect size measurement, which often contains temporal information such as "short-term ($\leq$3 months)" or "6-month follow-up."

D.3 LLM Extraction Approach

Follow-up periods were extracted using GPT-5 (OpenAI). For each outcome measurement, the model received the study characteristics and measurement label, and returned a follow-up period in days along with a justification. When the model could only determine a range (e.g., "3–6 months post-treatment"), it returned a range; when information was insufficient or contradictory, it returned null. When the model returned a range, we used the midpoint as the canonical value.

Study identifiers in the effect size data were matched to the characteristics database using fuzzy string matching (85% similarity threshold) to handle minor variations in study naming conventions across Cochrane reviews.

D.4 Extraction Prompt

The complete extraction prompt is reproduced below:

Task: Extract the follow-up period (days from end of treatment to outcome measurement) for a specific outcome measurement from a Cochrane meta-analysis.
Data Inputs:

- **Study characteristics:** Methods, participants, interventions (including timing/duration), outcomes (assessment timepoints), and notes from the "Characteristics of included studies" table.
- **Measurement label:** The outcome name, treatment comparison, and time window (e.g., "short-term $\leq$3 months").

Key Rules:
Measuring from End of Treatment (Critical): The follow-up period is measured from the **end of treatment**, not from randomization or baseline.

- **Default interpretation:** "6-month follow-up" or "assessed at 6 months" = 183 days (already post-treatment, no calculation needed).
- **When to calculate:** Only subtract treatment duration when explicit language indicates measurement from randomization, baseline, enrollment, or treatment start (e.g., "6 months from randomization" with an 8-week intervention = $183 - 56 = 127$ days).
- **Handling ambiguity:** If the reference point is unclear, assume measurement is from end of treatment.

Time Conversions: 1 week = 7 days; 1 month = 30.5 days; 1 year = 365 days.
When to Return Null: Return null when information is genuinely ambiguous, multiple time points are plausible, information is contradictory, or there is insufficient information. Do not guess.
[Study characteristics and measurement label inserted here]

D.5 Coverage and Distribution

Table D6 summarizes the coverage and distribution of extracted follow-up periods.

Figure D7 shows the distribution of follow-up periods in the analysis sample. The distribution is highly right-skewed: the median follow-up is approximately 1 month from the end of treatment, though the distribution extends to several years for some studies. The concentration of measurements near zero reflects that many studies assess outcomes at or shortly after treatment completion.

Approximately 66% of observations have a valid follow-up period. The 34% of observations with missing values typically correspond to cases where: (1) the outcome label contained no timing information (e.g., "Quality of life"), (2) study characteristics lacked explicit assessment timing, or (3) timing information was ambiguous or contradictory across sources.

D.6 Validation

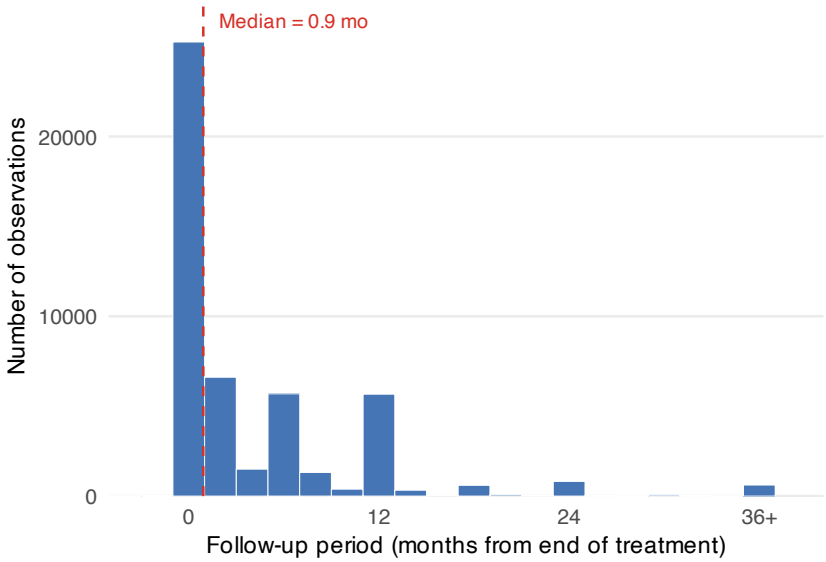
We validate the LLM-extracted follow-up periods by comparing them to values manually extracted from original papers by a research assistant. The validation sample consists of 25 randomly selected Cochrane reviews (678 studies, 5,127 outcome-level observations). For each observation, the research assistant read the original paper and recorded the follow-up period in days from end of treatment, following the same definition as the LLM extraction. Of the 5,127 observations in the

Table D6: *Follow-up Period Extraction Summary*

Panel A: Coverage by Source			
Source	Observations	Percent	
LLM extraction (scalar)	46,232	57.0%	
Missing (LLM returned null)	26,286	32.4%	
LLM extraction (range midpoint)	7,419	9.1%	
NA	1,185	1.5%	
Total	81,122	100.0%	
Panel B: Distribution of Follow-up Period			
Statistic	Days	Months	
Mean	134.9	4.4	
Median	28.0	0.9	
25th percentile	0.0	0.0	
75th percentile	183.0	6.0	
Min	0.0	0.0	
Max	10037.5	329.1	
SD	273.1	9.0	
Panel C: Coverage by Effect Measure Type			
Effect Measure	Total Obs.	With Follow-up	Coverage (%)
Standardised Mean Difference	34,105	23,525	69.0
Risk Ratio	20,139	13,241	65.7
Mean Difference	18,346	11,143	60.7
Odds Ratio	7,073	4,734	66.9
Risk Difference	453	319	70.4
NA	376	359	95.5
Hazard Ratio	204	127	62.3
Peto Odds Ratio	176	109	61.9
Rate Ratio	144	68	47.2

Notes: Panel A shows the source of follow-up period values. Panel B shows the distribution of follow-up periods for observations with non-negative values; 84 observations (0.16% of non-missing) yield negative follow-up periods, indicating outcomes measured during ongoing treatment rather than after its conclusion, and are excluded from Panel B and all temporal analyses. Panel C shows coverage rates by effect measure type. The 25th percentile of zero reflects end-of-treatment measurements; the right-skewed distribution reflects that most studies measure outcomes relatively soon after treatment ends, with fewer studies conducting long-term follow-up.

Figure D7: *Distribution of Follow-up Periods*



Notes: Histogram of follow-up periods (months from end of treatment) for the $N = 49{,}073$ observations with valid follow-up time data. Observations with follow-up > 36 months ($< 1\%$) are grouped in the rightmost bin. The dashed line marks the median.

sample, 906 have been coded to date (737 with a numeric value, 169 marked as indeterminate after reading the paper).

D.6.1 Accuracy of extracted values

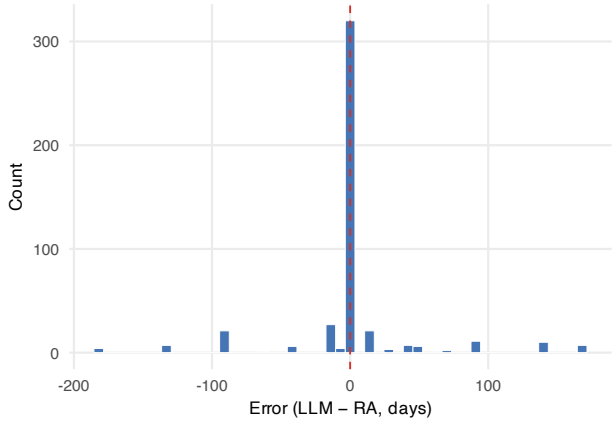
Figure D8 summarizes the accuracy of the LLM extraction for the 465 observations where both the LLM and the research assistant produced a numeric value. The left panel shows the distribution of errors (LLM − RA). The distribution is sharply peaked at zero, with a mean error of 3.1 days (median absolute error: 0 days). The right panel reports the fraction of observations for which the LLM value falls within a given tolerance of the human-coded value. The LLM and the research assistant agree exactly for 61.1% of observations, and 80.2% agree within ±30 days. The correlation between the two measures is 0.82.

D.6.2 Selection into observation

The LLM does not produce a follow-up value for every observation. In the validation sample, the LLM returns a value for 62% of observations (3,165 of 5,127). This missingness is not random: among observations for which the research assistant coded a follow-up period, those where the LLM also produced a value have longer average follow-up (mean difference 37.6 days, $p < 0.001$). The LLM is more likely to return missing when the true follow-up is zero (end-of-treatment measurement), where outcome labels often lack explicit timing information.

A separate concern is whether this selective observation biases the estimated fadeout slope. Table D9 compares inverse-variance weighted mean effect sizes between LLM-observed and LLM-missing observations within follow-up buckets defined by the research assistant's coding. At $t = 0$, where 59% of observations fall, there is no difference in mean effects ($p = 0.46$). At 0–3 months (17%

Figure D8: *Validation of LLM-Extracted Follow-up Periods*



(a) *Distribution of errors (LLM − RA)*

Tolerance	% within
Exact match	61.1%
± 7 days	68.8%
± 14 days	72.9%
± 30 days	80.2%
± 60 days	84.9%
± 90 days	85.6%

(b) *Cumulative accuracy*

Notes: Left panel: histogram of measurement error (LLM value minus research assistant value, in days) for the 465 observations where both produced a numeric follow-up period. Observations beyond the 1st and 99th percentile of the error distribution are trimmed for display. Right panel: cumulative fraction of observations for which the LLM-extracted value falls within the stated tolerance of the research assistant's value.

of observations), LLM-observed effects are larger than LLM-missing effects ($p = 0.006$), while at 3–12 months (24%) the pattern reverses at marginal significance ($p = 0.08$). These mixed patterns across horizons suggest that LLM missingness may generate bias in the estimated fadeout slope.

Table D9: *Effect Sizes by LLM Missingness Status and Follow-up Bucket*

		LLM observed		LLM missing			
Bucket	% of obs	n	IVW mean	n	IVW mean	Diff	p
$t = 0$	58.7	231	0.189	182	0.204	−0.015	0.461
0–3 months	16.8	84	0.312	34	0.187	0.126	0.006
3–12 months	24.3	142	0.177	29	0.239	−0.062	0.083

Notes: Inverse-variance weighted (IVW) mean effect sizes ($\hat{\theta} = \sum w_i Y_i / \sum w_i$, $w_i = 1/\text{se}_i^2$) for observations where the research assistant produced a numeric follow-up value, split by whether the LLM also produced a value. Follow-up buckets are defined by the research assistant's coding. "Diff" is LLM observed minus LLM missing; p-values are from a two-sided z-test using the IVW standard errors. One observation with follow-up > 12 months is omitted.

Future work will apply the correction methods proposed by Ludwig, Mullainathan, and Rambachan (2025) to formally adjust our fadeout estimates for this selective observation.

E Effect Size Standardization

To enable pooling across different effect size types, we convert all effects to a common scale of standard deviation (SD) units.

E.1 Difference Measures

Standardized Mean Differences (SMD): Used directly, as they are already expressed in SD units (e.g., Cohen's d, Hedges' g).

Mean Differences (MD): Converted to SD units by dividing by the best available estimate of the pooled standard deviation:

$$d = \frac{\text{MD}}{s_{\text{pooled}}} \tag{12}$$

where s_{pooled} is preferentially taken from (in order): the pooled SD reported in the study, the control group SD, the treatment group SD, or an estimated SD derived from other reported statistics.

E.2 Ratio Measures (Probit Transformation)

For binary outcomes reported as odds ratios (OR) or risk ratios (RR), we convert to standardized mean differences using the *probit transformation*, which computes the effect size directly from observed event rates (Glass et al., 1981):

$$d = \Phi^{-1}(p_e) - \Phi^{-1}(p_c) \tag{13}$$

where $p_e = n_e^{\text{events}}/n_e^{\text{total}}$ is the event rate in the treatment group, $p_c = n_c^{\text{events}}/n_c^{\text{total}}$ is the event rate in the control group, and $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative distribution function (probit function).

Interpretation. This "probit d" represents the difference in latent propensities between groups, assuming the binary outcome reflects an underlying continuous variable with a normal distribution and a fixed threshold. The transformation maps the proportion of events to its corresponding z-score on the standard normal distribution; the difference in z-scores is Cohen's d.

Advantages over log-odds transformation. This approach offers several advantages over the commonly-used Chinn (2000) log-odds transformation ($d = \log(\text{OR}) \times 0.55$):

- **No distributional assumption on ratios:** The probit method uses observed proportions directly, avoiding the assumption that the latent variable follows a logistic (rather than normal) distribution.
- **Works for all ratio types:** Risk ratios are converted identically to odds ratios—all via the same observed event rates—eliminating the need to assume RR $\approx$ OR (valid only for rare events). Hazard ratios (0.2% of observations) are excluded because converting survival-analysis measures to standardized mean differences requires assumptions about baseline hazard functions

and censoring distributions that cannot be verified from Cochrane summary statistics.

- **Consistency with Glass et al. tradition:** The probit d is the original formulation for effect size from 2×2 tables (Glass et al., 1981), predating the Chinn approximation.

Standard errors. We compute standard errors using the formula from Sánchez-Meca, Marín-Martínez, and Chacón-Moscoso (2003):

$$\text{Var}(d) = \frac{2\pi \cdot p_e(1-p_e) \cdot e^{z_e^2}}{n_e} + \frac{2\pi \cdot p_c(1-p_c) \cdot e^{z_c^2}}{n_c} \tag{14}$$

where $z_e = \Phi^{-1}(p_e)$ and $z_c = \Phi^{-1}(p_c)$. This variance formula accounts for the nonlinear transformation from proportions to the probit scale.

Boundary handling. When $p = 0$ or $p = 1$, the probit function is undefined ($\pm\infty$). We apply continuity correction by bounding proportions to the interval $[0.001, 0.999]$ before transformation, consistent with standard practice.

Fallback: Chinn transformation. When the Cochrane data do not include the underlying event counts (approximately 18% of ratio-based observations), we fall back to the Chinn (2000) transformation:

$$d = \log(\text{OR}) \times \frac{\sqrt{3}}{\pi} \approx \log(\text{OR}) \times 0.5513 \tag{15}$$

with standard error $\text{SE}_d = \text{SE}_{\log(\text{OR})} \times 0.5513$. This transformation assumes a logistic latent distribution.

Coverage. In our final dataset, 78% of ratio-based observations use the probit transformation (event counts available), 18% use the Chinn fallback, and 4% have extreme proportions ($p = 0$ or $p = 1$ in one or both groups) where neither method is applicable. The `conversion_method` variable in the data tracks which method was used for each observation.

F Model Specification

This appendix provides complete specifications for the Bayesian hierarchical models used in our analysis. All models are estimated using Markov chain Monte Carlo (MCMC) via Stan (Carpenter et al., 2017).

F.1 Analysis-Level Model

For each analysis j, we estimate a Bayesian hierarchical model with a common linear time trend across studies:

Likelihood.

$$Y_{ijt} \sim \mathcal{N}(\theta_{ij} + \beta_j \cdot t_{ijt}, \sigma^2_{ijt}) \tag{16}$$

where Y_{ijt} is the effect size for study i in analysis j at follow-up time t (in months), expressed in SD units after the transformations described in Appendix E. The standard error σ_{ijt} is treated as known from the original study.

Study-level heterogeneity.

$$\theta_{ij} \sim \mathcal{N}(\alpha_j, \tau^2_j) \tag{17}$$

where θ_{ij} is a study-specific intercept (capturing baseline heterogeneity in effect magnitude), α_j is the population mean intercept for analysis j, and τ_j is between-study heterogeneity.

Parameters of interest. The key parameter is β_j, the fadeout rate common to all studies within the analysis. A negative β_j indicates declining effects over time (fadeout); a positive β_j indicates growing effects.

Priors.

$$\alpha_j \sim \mathcal{N}(0, 2^2) \tag{18}$$

$$\beta_j \sim \mathcal{N}(0, 0.5^2) \tag{19}$$

$$\tau_j \sim \text{Half-Cauchy}(0, 0.5) \tag{20}$$

Prior rationale.

- **Intercept prior** (α_j): The $\mathcal{N}(0, 4)$ prior (95% interval: $[-4, 4]$ SD) is weakly informative, accommodating effect sizes from small to very large while placing most prior mass in the range of typical intervention effects.
- **Slope prior** (β_j): The $\mathcal{N}(0, 0.5^2)$ prior assigns 95% probability to slopes between -1 and $+1$ SD per month, corresponding to annual changes of ± 12 SD—far exceeding any plausible fadeout rate. This prior is deliberately diffuse to allow the data to dominate.

- **Heterogeneity prior** (τ_j): The Half-Cauchy(0, 0.5) prior is recommended by Gelman (2006) as a weakly informative prior for variance components in hierarchical models. It allows substantial heterogeneity while regularizing toward zero when data are sparse.

MCMC settings. Each analysis-level model is estimated with 4 chains, 10,000 iterations per chain (5,000 warmup), target acceptance rate $\delta = 0.99$, and maximum tree depth 15. These conservative settings ensure reliable convergence across the heterogeneous set of analyses.

F.2 Static Model (Comparison)

For comparison with the dynamic model, we also estimate a static (intercept-only) model that ignores follow-up timing:

$$Y_{ijt} \sim \mathcal{N}(\theta_{ij}, \sigma^2_{ijt}) \tag{21}$$

$$\theta_{ij} \sim \mathcal{N}(\alpha_j, \tau^2_j) \tag{22}$$

with the same priors on α_j and τ_j. This model pools all effect sizes regardless of when outcomes were measured, corresponding to standard random-effects meta-analysis. Comparing static and dynamic model predictions allows us to assess whether accounting for effect dynamics changes substantive conclusions (Section 3.1).

F.3 Aggregate Model (Meta-Meta-Analysis)

To estimate the overall average fadeout rate across all analyses, we fit a hierarchical model pooling observations from all analyses. A key feature of our model is *crossed study random effects*: the same study often appears in multiple analyses (measuring different outcomes), and we account for this non-independence by including study-specific random effects that cross analysis boundaries.

Likelihood.

$$Y_{ijt} \sim \mathcal{N}(\theta_i + \alpha_j + \beta_j \cdot t_{ijt}, \sigma^2_{ijt}) \tag{23}$$

where θ_i is a study-specific random effect that applies to *all* observations from study i, regardless of which analysis they belong to, and α_j, β_j are analysis-specific intercepts and slopes.

Crossed study random effects. The study effect θ_i captures study-level heterogeneity (e.g., sample quality, implementation fidelity, context) that affects all outcomes measured in that study:

$$\theta_i \sim \mathcal{N}(0, \tau^2_{\text{study}}) \tag{24}$$

This crossed structure is important because approximately 80% of studies appear in two or more analyses (measuring different outcomes), with the average study appearing in over six analyses. Without crossed effects, we would treat these observations as independent, underestimating standard errors.

Correlated analysis-level random effects. We model analysis-specific intercepts and slopes as draws from a bivariate normal distribution:

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma \right) \tag{25}$$

where

$$\Sigma = \begin{pmatrix} \tau_\alpha^2 & \rho\,\tau_\alpha\tau_\beta \\ \rho\,\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix} \tag{26}$$

The correlation ρ captures the relationship between initial effect magnitude and fadeout rate. A negative ρ indicates that larger initial effects fade faster (consistent with mean reversion); a positive ρ indicates larger effects persist longer.

Hyperpriors.

$$\mu_\alpha \sim \mathcal{N}(0.3, 0.5^2) \tag{27}$$

$$\mu_\beta \sim \mathcal{N}(0, 0.02^2) \tag{28}$$

$$\tau_\alpha \sim \text{Half-Cauchy}(0, 0.5) \tag{29}$$

$$\tau_\beta \sim \text{Half-Cauchy}(0, 0.02) \tag{30}$$

$$\tau_{\text{study}} \sim \text{Half-Cauchy}(0, 0.5) \tag{31}$$

$$\rho \sim \text{Uniform}(-1, 1) \tag{32}$$

Hyperprior rationale.

- **Mean intercept** (μ_α): The prior $\mathcal{N}(0.3, 0.5^2)$ is centered on a small-to-medium effect size, reflecting the typical magnitude of psychosocial interventions in the meta-analytic literature. The 95% prior interval (-0.7 to 1.3 SD) spans the full range of plausible average effects.
- **Mean slope** (μ_β): The prior $\mathcal{N}(0, 0.02^2)$ places 95% probability on average fadeout rates between -0.04 and $+0.04$ SD per month, corresponding to annual fadeout of ± 0.48 SD. This represents substantial fadeout—an intervention with initial effect of 0.3 SD would decline to zero within 8 months at the prior's 97.5th percentile rate. The prior is thus conservative in that it allows for substantial fadeout while remaining centered at zero (no prior expectation of fadeout vs. persistence).
- **Between-analysis heterogeneity** (τ_α, τ_β): Half-Cauchy priors with scales calibrated to the expected variation in intercepts (0.5 SD) and slopes (0.02 SD/month). These allow substantial heterogeneity while providing regularization.
- **Study heterogeneity** (τ_{study}): Half-Cauchy(0, 0.5) prior, matching the scale for analysis intercepts. This allows the data to determine how much variance is at the study level vs. the analysis level.

- **Correlation** (ρ): A uniform prior on $(-1, 1)$ is uninformative, allowing the data to determine whether intercepts and slopes are correlated.

Variance partitioning. The model partitions heterogeneity into study-level (τ^2_{study}) and analysis-level (τ^2_α) components. The intraclass correlation coefficient (ICC) for studies is:

$$\text{ICC}_{\text{study}} = \frac{\tau^2_{\text{study}}}{\tau^2_{\text{study}} + \tau^2_\alpha} \tag{33}$$

This measures what fraction of the unexplained heterogeneity in effect sizes is due to study-level factors (which apply across all outcomes from that study) versus analysis-level factors (which are specific to particular outcomes).

Parameterization. To improve MCMC efficiency, we use a non-centered parameterization. Analysis-specific effects are constructed using Cholesky decomposition:

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix} + \mathbf{DLz}_j \tag{34}$$

where $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ are standard normal deviates, $\mathbf{D} = \text{diag}(\tau_\alpha, \tau_\beta)$, and $\mathbf{L}$ is the Cholesky factor of the correlation matrix. Study effects use a similar non-centered form: $\theta_i = \tau_{\text{study}} \cdot z_i$ where $z_i \sim \mathcal{N}(0, 1)$.

MCMC settings. The aggregate model is estimated with 4 chains, 4,000 iterations per chain (2,000 warmup), target acceptance rate $\delta = 0.95$, and maximum tree depth 12. Fewer iterations are needed because the model pools information across many analyses.

F.4 Convergence Diagnostics

We assess convergence using standard criteria:

- $\hat{R}$ **(potential scale reduction factor):** All parameters must have $\hat{R} < 1.01$, indicating chains have mixed well.
- **Effective sample size (ESS):** We require minimum ESS > 400 for reliable posterior summaries, though most analyses achieve ESS > 1000.
- **Divergent transitions:** We monitor for divergent transitions, which indicate the sampler encountered regions of high curvature. The high target acceptance rate ($\delta = 0.95$–0.99) reduces divergences; analyses with persistent divergences are flagged for inspection.
- **Evidence of dynamics:** We define an analysis as showing "evidence of dynamics" if the posterior probability of direction exceeds 0.95—that is, if $\max\{P(\beta_j > 0 \mid \text{data}), P(\beta_j < 0 \mid \text{data})\} > 0.95$. This threshold corresponds to the 90% credible interval excluding zero.

Across the 1,503 analysis-level models, the median $\hat{R}$ for the slope parameter is 1.000 (IQR: 1.000–

1.001), and 99.4% of analyses have $\hat{R} < 1.01$. The aggregate model with crossed study random effects achieves $\hat{R} < 1.02$ for all parameters with minimum ESS > 350.

F.5 Direction Standardization

Treatment effect directions are inherently arbitrary: an intervention that "reduces depression symptoms" and one that "increases psychological well-being" may capture the same underlying improvement, but with opposite-signed effect sizes. To ensure that fadeout is consistently interpretable as effects declining toward zero, we standardize directions within each analysis.

F.5.1 Procedure

For each analysis j, we compute the inverse-variance weighted mean of all effect sizes:

$$\bar{Y}_j^w = \frac{\sum_i w_i Y_{ij}}{\sum_i w_i}, \quad \text{where } w_i = \frac{1}{\text{SE}_{ij}^2} \tag{35}$$

The *flip indicator* s_j is then:

$$s_j = \begin{cases} -1 & \text{if } \bar{Y}_j^w < 0 \\ +1 & \text{otherwise} \end{cases} \tag{36}$$

We define the *direction-standardized effect size* as $Y_{ij}^{\text{flip}} = s_j \cdot Y_{ij}$.

For ratio measures (odds ratios, risk ratios), the procedure is applied after converting to SD units via the probit transformation (Appendix E.2). On the SD scale, flipping simply negates the effect size, as with difference measures.

F.5.2 Rationale for Inverse-Variance Weighting

We use inverse-variance weighting rather than simple averaging when computing $\bar{Y}_j^w$ because:

1. **Precision:** More precisely estimated effects (smaller SE) receive more weight, reducing the influence of noisy estimates on the direction determination.
2. **Consistency with meta-analysis:** Inverse-variance weighting is the standard approach in meta-analysis, so our direction determination aligns with how the effects will be aggregated.
3. **Robustness to outliers:** Imprecise outliers with extreme values but large standard errors have minimal influence on the weighted mean.

F.5.3 Interpretation

After direction standardization, a negative slope $\beta_j < 0$ indicates that treatment effects are declining toward zero over time (fadeout), while a positive slope $\beta_j > 0$ indicates effects are growing (moving further from zero). This interpretation holds regardless of how the original outcome was coded.

Importantly, direction standardization is a sign convention, not a substantive claim that the intervention is beneficial. If an intervention is on average harmful—and flipping codes that harm as a positive effect—then fadeout ($\beta_j < 0$) correctly indicates that harm is dissipating over time. The slope's sign always reflects the direction of change relative to zero, not relative to any normative

benchmark. In particular, because the slope β_j is defined on the same scale before and after flipping (with only the sign of the intercept changed), the *magnitude and direction* of the slope are unaffected by the choice of sign convention: flipping reverses both the intercept and the slope, so the trajectory's shape is preserved.

The flip indicator s_j is retained in the dataset, allowing researchers to identify which analyses were flipped and to recover the original effect direction if needed.

G Analysis Definition

An *analysis* pools effect sizes for the same intervention-outcome pair measured at different follow-up periods. This section describes how we define analyses and handle edge cases that arise when grouping effect sizes from Cochrane systematic reviews.

G.1 Basic Construction

Each analysis is defined by a unique combination of three elements:

1. **Review ID:** The Cochrane review identifier (e.g., CD001234), which defines the intervention domain
2. **Cleaned outcome label:** The outcome measure with temporal information removed (see below)
3. **Comparison label:** The specific treatment-versus-control comparison within the review

The analysis identifier is constructed as: review × outcome label × comparison label.

G.2 Outcome Label Standardization

Cochrane reviews often include temporal information in outcome labels (e.g., "Depression at 6 months," "Anxiety (short-term)," "Quality of life - 1 year follow-up"). Since our goal is to estimate how effects change over time *within* each analysis, we strip temporal markers from outcome labels so that measurements of the same construct at different times are grouped together. The follow-up period is coded separately as t_{ijt} and enters the model as a covariate.

Temporal markers include explicit time points ("at 3 months," "1-year"), relative descriptors ("short-term," "long-term," "post-intervention"), and implicit markers ("baseline," "endpoint," "follow-up"). After stripping these markers, "Depression at 6 months" and "Depression at 2 years" become a single outcome "Depression" measured at two different follow-up times.

G.3 Effect Type Separation

When a review-outcome-comparison combination contains both continuous measures (standardized and unstandardized mean differences) and ratio measures (odds ratios, risk ratios), we separate these into distinct analyses by appending the effect type to the analysis identifier.

G.4 Subscale Handling

Some reviews use subgroup labels to distinguish conceptually distinct outcomes (e.g., PANSS positive versus negative symptoms) rather than time periods. When subgroup labels represent distinct subscales or instruments, we include the subscale identifier in the analysis definition to prevent inappropriate pooling.

H Power Simulation

We assess statistical power using Monte Carlo simulation calibrated to the empirical characteristics of our high-quality analyses. For each of 163 analyses meeting quality criteria (at least ten studies and interquartile range of follow-up times ≥ 3 months), we:

1. **Extract empirical parameters:** number of studies, proportion of panel studies, follow-up time range, and the distribution of standard errors.
2. **Generate synthetic data:** For each simulation, we generate study-level effects from $\theta_i \sim N(0, \tau^2)$ where $\tau \in \{0.15, 0.30, 0.50\}$ captures between-study heterogeneity. Follow-up times are drawn uniformly over the analysis's observed range. Observations are generated as $Y_{it} = \theta_i + \beta \cdot t + \epsilon_{it}$ where β is the true slope and $\epsilon_{it} \sim N(0, \sigma_{it}^2)$ with standard errors drawn from the analysis's empirical distribution.
3. **Fit the model:** We fit our Bayesian hierarchical model (Section 2.3) to each simulated dataset and record whether the 90% credible interval for the slope excludes zero.
4. **Compute power:** Power is the proportion of simulations (out of 50 per configuration) in which the credible interval excludes zero. The minimum detectable effect (MDE) is the smallest true slope β at which power reaches 80%.

We simulate across true slopes $\beta \in \{0, 0.01, 0.02, 0.03, 0.05, 0.08\}$ SD per month. The case $\beta = 0$ provides the false positive rate, which averages 4.2% across configurations—close to the nominal 5% level implied by the 90% credible interval threshold.

I IPW Selection Correction Methodology

We use inverse probability weighting (IPW) to correct for potential selection into long-run follow-up. The approach follows standard attrition correction: observations with long-run data are the "non-attriters," and we reweight them to represent the full population (including "attriters" without long-run data). We implement corrections at two levels: analysis-level and study-level.

I.1 Analysis-Level IPW

I.1.1 Propensity Model

We estimate the probability that an analysis has any long-run follow-up (>6 months) using logistic regression. The propensity model includes:

- **Short-run effect characteristics:** Mean effect size for observations with $t \leq 6$ months (direction-standardized), and mean standard error of these estimates.
- **Analysis size:** Log number of short-run observations and log total sample size.
- **Effect measure type:** Binary indicator for originally ratio-based measures (OR, RR) versus difference measures (SMD, MD).
- **Intervention type:** Binary indicators for each of ten PICO intervention clusters.

The propensity score is $\hat{p}_j = P(\text{long-run} \mid X_j)$.

I.1.2 Weight Construction

We compute stabilized IPW weights following the standard attrition correction formula:

$$w_j = \frac{P(\text{long-run})}{P(\text{long-run} \mid X_j)} = \frac{\bar{p}}{\hat{p}_j} \tag{37}$$

where $\bar{p}$ is the marginal (unconditional) probability of having long-run follow-up, and $\hat{p}_j$ is the propensity score. Analyses with characteristics *atypical* of long-run analyses—those with low $\hat{p}_j$ despite having long-run data—receive higher weights. This reweights the long-run sample to match the covariate distribution of the full population. Weights are trimmed at the 95th percentile to limit the influence of extreme values.

I.2 Study-Level IPW

I.2.1 Definition of Long-run

At the study level, "long-run" is defined relative to the analysis's temporal distribution. Within each analysis j, we compute the median follow-up time $\tilde{t}_j$. A study has long-run follow-up if it has at least one observation with $t > \tilde{t}_j$.

I.2.2 Propensity Model

We estimate the probability that a study has long-run follow-up as a function of study-level characteristics:

- **Short-run effect size:** Mean effect size for the study's short-run observations ($t \leq \tilde{t}_j$).

- **Precision:** Mean standard error of the study's short-run observations.
- **Sample size:** Total sample size of the study.
- **Analysis characteristics:** Effect measure type and intervention cluster indicators.

I.2.3 Weight Construction

Stabilized weights are computed analogously:

$$w_i = \frac{P(\text{long-run study})}{P(\text{long-run study} \mid X_i)} \tag{38}$$

where the marginal probability is computed across all studies and the propensity score is study-specific. Studies with characteristics atypical of long-run studies receive higher weights.

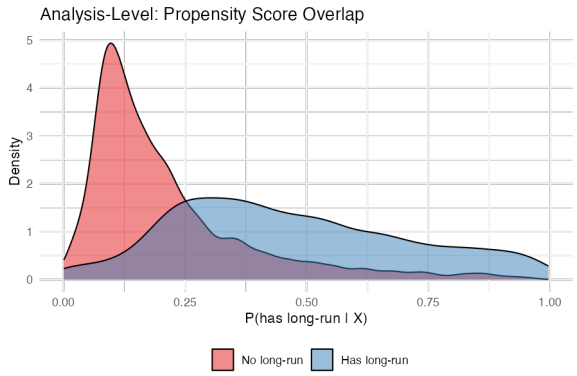
I.3 Diagnostics

Figure I10 presents diagnostic checks for the analysis-level IPW. Panel (a) shows propensity score distributions for analyses with and without long-run follow-up. The distributions substantially overlap, with few analyses having extreme propensity scores (<0.1 or >0.9), indicating that IPW is well-supported by common support. Panel (b) shows the distribution of trimmed IPW weights, which are moderately variable but without extreme outliers after trimming.

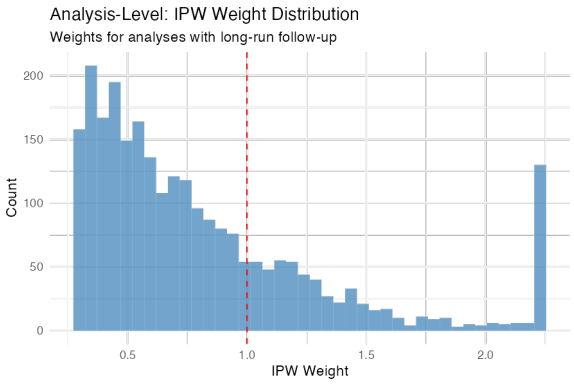
Figure I11 presents analogous diagnostics for the study-level IPW. Propensity scores again show substantial overlap between studies with and without long-run follow-up, and the weight distribution is well-behaved after trimming.

Figure I10: *IPW Diagnostics (Analysis Level)*

(a) Propensity Score Overlap



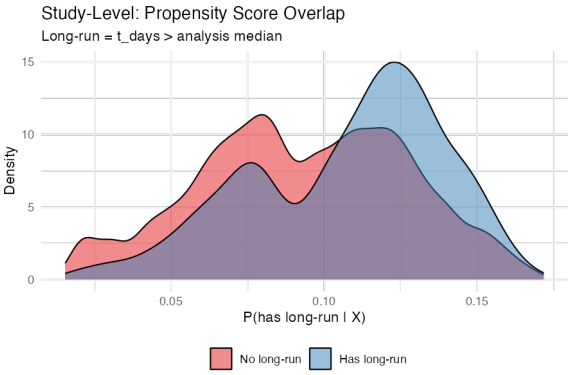
(b) IPW Weight Distribution



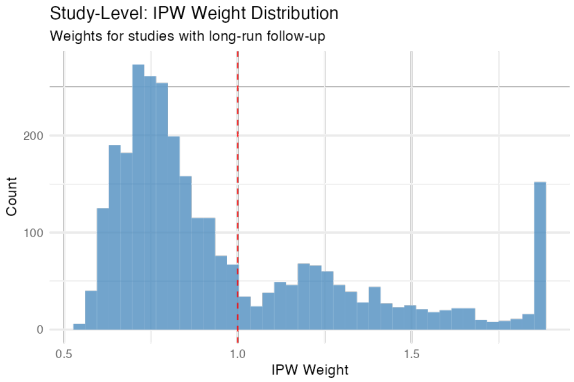
Notes: Panel (a) shows the distribution of propensity scores (predicted probability of having long-run follow-up) for analyses with and without long-run data. Substantial overlap indicates IPW is well-supported by common support. Panel (b) shows the distribution of stabilized IPW weights across analyses with long-run follow-up. Weights are computed as $w_j = P(\text{long-run})/P(\text{long-run} \mid X_j)$; analyses with characteristics atypical of long-run analyses (low propensity) receive higher weights. Weights trimmed at 95th percentile.

Figure I11: *IPW Diagnostics (Study Level)*

(a) Propensity Score Overlap



(b) IPW Weight Distribution



Notes: Panel (a) shows the distribution of propensity scores (predicted probability of having long-run follow-up) for studies with and without long-run data, where "long-run" is defined as having observations beyond the analysis median follow-up time. Panel (b) shows the distribution of stabilized IPW weights across studies with long-run follow-up. Weights trimmed at 95th percentile.

J Pessimistic Selection Bounds

We construct pessimistic bounds on fadeout by augmenting the data with synthetic observations that assume complete effect disappearance for studies with missing long-run data. Two bounding approaches are implemented:

J.1 Fixed Bound (Y=0 at 36 Months)

1. **Identify studies with missing long-run data**: For each study, we check whether it has any observations beyond 24 months post-treatment. Studies without such observations are classified as having "missing" long-run data.
2. **Add synthetic observations**: For each study with missing long-run data, we add a single synthetic observation at $t = 36$ months with:
   - Effect size $Y = 0$ (complete fadeout to null effect)
   - Standard error equal to the SE from that study's latest actual observation
3. **Re-estimate the model**: We fit the same Bayesian hierarchical model (Section 2.4) to the augmented dataset, yielding bounded estimates of the population-mean fadeout rate μ_β.

This approach assumes the worst case for studies with missing long-run data while leaving studies with observed long-run data unchanged.

J.2 Relative Bound (Y=0 at $1.2 \times t_{\max}$)

1. **Calculate analysis-specific time horizons**: For each analysis j, we compute $t_j^* = 1.2 \times \max(t_{ij})$—that is, 20% beyond the maximum observed follow-up time in that analysis.
2. **Add synthetic observations for all studies**: Every study within each analysis receives a synthetic observation at t_j^* with effect size $Y = 0$ and SE equal to the study's latest observed SE.
3. **Re-estimate the model**: As above.

This more aggressive bound assumes that *all* studies would show complete fadeout shortly after their last observation, regardless of whether they have long-run data. It is deliberately extreme and likely overstates true fadeout.

J.3 Results Summary

Table J12 summarizes the bounded estimates compared to the original specification.

Table J12: *Pessimistic Bounds on Fadeout Rate*

Specification	Slope (SD/month)	95% CI	Annual Fadeout
Original	−0.0003	[−0.0035, 0.0028]	−0.00 SD/year
Type 1 (Fixed)	−0.0071	[−0.0075, −0.0067]	−0.09 SD/year (23.9×)
Type 2 (Relative)	−0.0209	[−0.0223, −0.0195]	−0.25 SD/year (70.4×)

Notes: Multiplier (in parentheses) indicates the ratio of bounded to original fadeout rate. The 36-month bound adds synthetic Y=0 at 36 months for studies without >24 month data. The $1.2 \times t_{\max}$ bound adds synthetic Y=0 at 1.2× max observed time for all studies.

K Robustness to Functional Form

Our main specification models fadeout as linear in follow-up time. This section tests robustness to two non-linear alternatives—a piecewise linear spline and an exponential decay model—at both the individual analysis level and the aggregate meta-meta-analysis level.

K.1 Alternative Model Specifications

Piecewise linear spline. We replace the single linear slope with a piecewise linear function that can change gradient at pre-specified knots. For the analysis-level models, knots are placed adaptively at the 25th, 50th, and 75th percentiles of each analysis's time distribution. For the meta-meta-analysis, knots are fixed at 3, 6, and 12 months post-treatment. The spline is parameterized as:

$$Y_{ij}(t) = \alpha_{ij} + \beta_1 t + \beta_2(t-k_1)_+ + \beta_3(t-k_2)_+ + \beta_4(t-k_3)_+$$

where $(x)_+ = \max(0, x)$ and k_1, k_2, k_3 are the knot positions. This maintains a continuous trajectory while allowing the slope to change at each knot. The coefficients $\beta_2, \beta_3, \beta_4$ represent slope *changes* at each knot—if these are indistinguishable from zero, the data do not support non-linear fadeout.

The analysis-level spline model uses the same hierarchical structure as the linear model: study-level random intercepts with a half-Cauchy prior on the heterogeneity parameter (τ), and weakly informative priors on the intercept ($\mathcal{N}(0, 2)$) and slope coefficients ($\mathcal{N}(0, 0.5)$). At the meta-meta level, the spline model adds analysis-level random effects on the intercept and each slope coefficient, with half-Cauchy priors on the corresponding heterogeneity parameters.

Exponential decay. We model the effect trajectory as:

$$Y_{ij}(t) = \alpha_{ij} \exp(-\lambda_{ij}\, t) + c_{ij}$$

where α_{ij} is the initial "excess" effect above a long-run asymptote c_{ij}, and $\lambda_{ij} > 0$ is the decay rate. The initial effect is $\alpha_{ij} + c_{ij}$, and the half-life of the excess component is $\ln(2)/\lambda_{ij}$. This specification allows effects to stabilize at a non-zero level rather than declining indefinitely toward zero.

The analysis-level exponential model uses weakly informative priors: $\alpha \sim \mathcal{N}(0.2, 0.5)$, $\lambda \sim$ Half-Normal(0.1), $c \sim \mathcal{N}(0.1, 0.3)$, with a half-Cauchy prior on the study-level heterogeneity. At the meta-meta level, analysis-level random effects are placed on the log-decay rate $\log(\lambda)$ to ensure positivity, with normal priors on the population mean ($\mu_{\log\lambda} \sim \mathcal{N}(-2, 1)$).

All models are estimated using Hamiltonian Monte Carlo via Stan, with 4 chains of 10,000 iterations (5,000 warmup) for analysis-level models and 4,000 iterations (2,000 warmup) for the meta-meta models. The same MCMC diagnostics (divergent transitions, $\hat{R}$, effective sample size) are monitored as for the linear specification.

K.2 Analysis-Level Results

Each non-linear model is fit to the same 1,503 qualifying analyses used in the linear batch (selection criteria: time IQR $\geq$ 3 months, $\geq$ 3 studies per analysis). For each analysis, we compare the estimated fadeout rate across models: the linear slope (SD/month), the spline average slope (SD/month, averaging across segments weighted by segment duration), and the exponential implied fadeout rate (annual fadeout / 12, for comparability).

Table K13 summarizes the distribution of analysis-level slopes across all three model specifications.

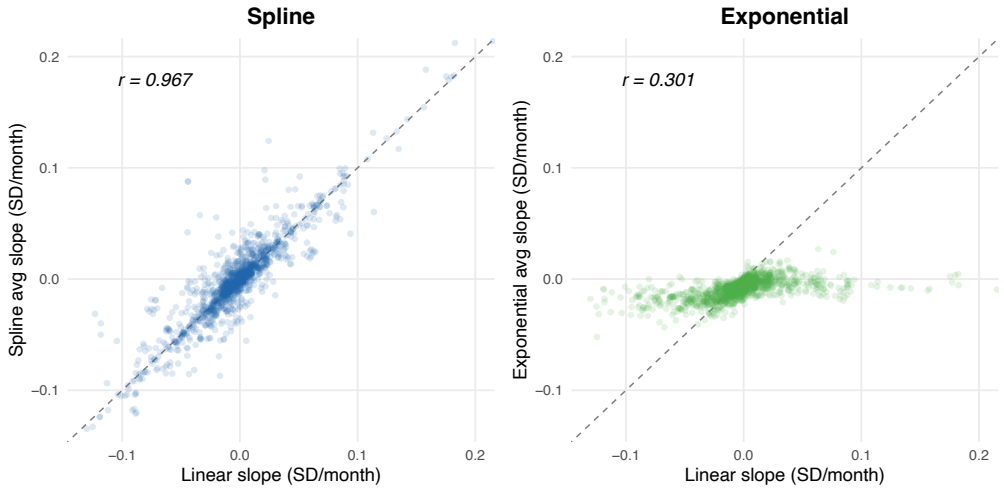
Table K13: *Analysis-Level Slope Estimates Across Model Specifications*

	Linear	Spline	Exponential
N analyses	1503	1503	1503
Mean slope (SD/month)	-0.0007	0.0004	-0.0086
Median slope (SD/month)	-0.0020	-0.0015	-0.0078
SD of slopes	0.0652	0.0666	0.0088
% negative slopes	56.4	54.2	86.7
Correlation with linear	—	0.967	0.301
% with non-linearity evidence	—	1.3	—

Notes: Summary statistics computed over 1,503 analyses that meet the inclusion criteria (time IQR $\geq$ 3 months, $\geq$ 3 studies). All slopes are expressed in SD per month. Linear slopes are converted from per-day estimates ($\times$30); exponential slopes are converted from annual fadeout ($\div$12). Correlation is the Pearson correlation between each non-linear model's slope and the linear slope. Non-linearity evidence for the spline model indicates the percentage of analyses where at least one slope-change coefficient's 95% credible interval excludes zero.

Figure K14 plots the analysis-level slope from the linear model against the corresponding slope from each non-linear model. Points cluster tightly around the 45-degree line, confirming that the three specifications produce concordant estimates at the individual analysis level.

Figure K14: *Analysis-Level Slope Comparison: Linear vs Non-Linear Models*

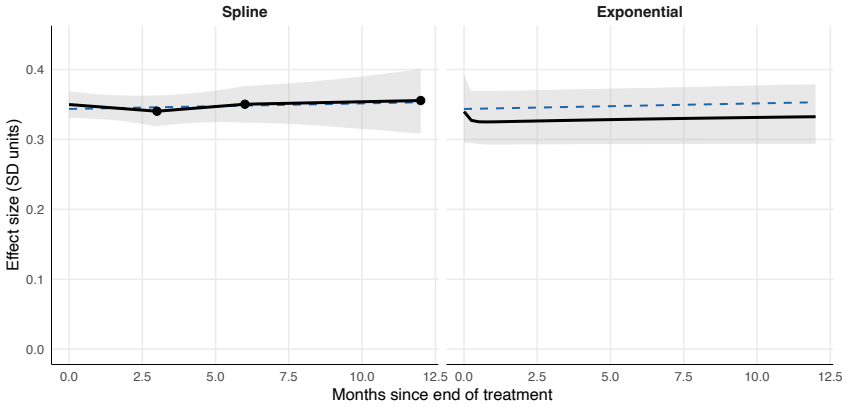


Notes: Each point represents one of 1,503 analyses. The x-axis shows the linear model's estimated slope (SD/month) and the y-axis shows the corresponding non-linear model's slope. Left panel: spline average slope. Right panel: exponential implied monthly fadeout. The dashed line is the 45-degree line (perfect agreement). Pearson correlations are reported in each panel.

K.3 Meta-Meta-Analysis Results

At the aggregate level, the spline and exponential models are estimated on the full pooled dataset using the same crossed random effects structure as the linear meta-meta-analysis (Appendix F). Figure K15 compares the predicted fadeout curves from all three models.

Figure K15: *Fadeout Curve Comparison: Linear vs Non-Linear Models*



Notes: Each panel compares a non-linear model (solid line) against the linear baseline (dashed line). Left: piecewise linear spline with knots at 3, 6, and 12 months (dots). Right: exponential decay $Y(t) = \mu_\alpha e^{-\mu_\lambda t} + \mu_c$. Shaded regions show 95% credible intervals for the non-linear model. The linear model is $Y(t) = \mu_\alpha + \mu_\beta t$. All models estimated on the same sample using Bayesian hierarchical methods with crossed study random effects.

Table K16 reports the key parameter estimates and predicted effects at standard time horizons. All three models agree closely on the initial effect size (≈ 0.34 SD) and show minimal fadeout over the first two years. The exponential model's asymptote and half-life estimates are reported as raw posterior summaries; the very large half-life reflects the near-absence of decay in the data.

Table K16: *Meta-Meta-Analysis Parameter Estimates by Model*

Parameter	Linear	Spline	Exponential
Initial effect (SD)	0.344 [0.326, 0.361]	0.350 [0.331, 0.369]	0.340 [0.295, 0.393]
Effect at 12mo (SD)	0.353 [0.311, 0.395]	0.356 [0.308, 0.401]	0.332 [0.294, 0.379]
Effect at 24mo (SD)	0.363 [0.279, 0.446]	0.389 [0.287, 0.490]	0.337 [0.294, 0.388]
Annual fadeout (SD/yr)	0.010 [-0.034, 0.054]	0.006 [-0.043, 0.053]	0.007 [-0.036, 0.080]
Asymptote (SD)	—	—	3.755 [0.296, 14.351]
Half-life (months)	—	—	66627.516 [0.062, 410123.098]

Notes: Point estimates are posterior means with 95% credible intervals in brackets. Initial effect is the predicted effect at $t = 0$. Annual fadeout is the predicted change in effect size over 12 months. Asymptote and half-life are specific to the exponential model.

Table K17 presents the spline model's segment-specific slopes. None of the four segments shows a slope that is statistically distinguishable from zero, and the slope-change parameters ($\beta_2, \beta_3, \beta_4$) all have 95% credible intervals that comfortably include zero. This provides no evidence that the fadeout rate varies across time segments.

Table K17: *Spline Model: Segment-Specific Slopes*

Segment	Slope (SD/month)	95% CI
0–3 months	-0.0032	[-0.0096, 0.0032]
3–6 months	0.0033	[-0.0035, 0.0101]
6–12 months	0.0009	[-0.0044, 0.0061]
12+ months	0.0028	[-0.0029, 0.0084]

Notes: Posterior mean slopes for each time segment from the piecewise linear spline meta-meta-analysis. Knots are placed at 3, 6, and 12 months. The 95% credible intervals for all segments include zero, indicating no statistically significant fadeout in any period.

L Robustness to Within-Study Correlation

For panel studies—studies that measure outcomes at multiple follow-up periods on the same participants—effect estimates at different time points are correlated. Our baseline model (Section 2) includes study-specific intercepts θ_{ij} that absorb cross-study differences in effect levels, but uses a common slope β_j across all studies within an analysis. This means that multiple time points from one panel study are treated as independent observations informing the slope, conditional on the study intercept and the linear trend. If within-study residuals are positively correlated (as they would be when the same participants contribute to each time point), this overstates the effective information content of panel observations and may understate uncertainty about the slope.

To assess the sensitivity of our results to this assumption, we re-estimate the analysis-level model with study-specific random slopes:

$$Y_{ijt} \sim N(\theta_{ij} + \beta_{ij} \cdot t_{ijt}, \sigma^2_{ijt}) \tag{39}$$

$$\theta_{ij} \sim N(\alpha_j, \tau^2_j) \tag{40}$$

$$\beta_{ij} \sim N(\beta_j, \tau^2_\beta) \tag{41}$$

where β_{ij} is now a study-specific slope drawn from a population distribution centered on β_j with between-study standard deviation τ_β. This model treats each panel study's time series as providing one noisy estimate of the slope, rather than multiple independent observations. Even for studies with a single time point, the random slope is estimated (with high shrinkage toward β_j) via partial pooling.

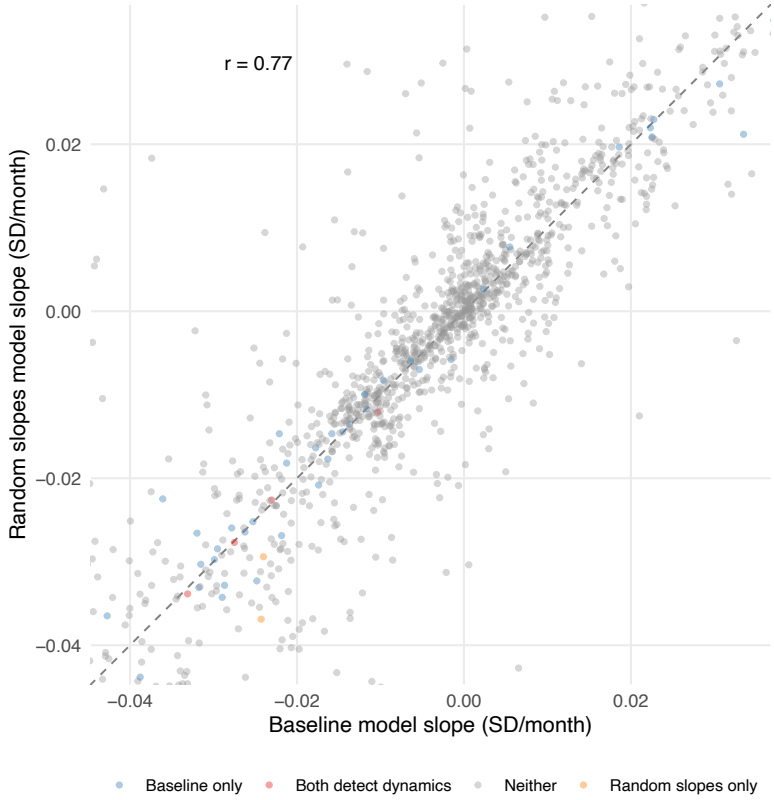
We place a half-Cauchy(0, 0.1) prior on τ_β, reflecting the expectation that study-level slope heterogeneity should be modest relative to the population slope. The model uses a non-centered parameterization for both random intercepts and slopes to improve sampling efficiency. MCMC settings are slightly more conservative than the baseline (12,000 iterations, 6,000 warmup, $\delta = 0.995$) to accommodate the additional parameters.

Figure L18 compares the population slope estimates from the baseline (common slope) and random slopes models across 1,484 matched analyses. The two models yield highly correlated slope estimates ($r = 0.77$), with nearly identical means (−0.0010 vs −0.0011 SD/month). As expected, the random slopes model produces wider credible intervals—the median CI width is 1.6 times larger than the baseline—reflecting the additional uncertainty from estimating study-specific slopes. Consequently, the number of analyses with detectable dynamics (posterior probability of direction > 0.95) falls from 80 under the baseline to 14 under random slopes (0.9%). Of these, 11 are detected by both models. The estimated between-study slope heterogeneity is modest: the median $\hat{\tau}_\beta$ is 0.046 SD/month, small relative to the population slope uncertainty.

The key conclusion is unchanged: relaxing the common-slope assumption does not reveal hidden fadeout. Point estimates remain near zero, and the wider credible intervals merely reduce statistical power to detect dynamics. This is consistent with the power analysis in Section 5.1: in the presence

of moderate within-study correlation, the effective sample size per panel study is smaller, making it harder to distinguish small slopes from noise.

Figure L18: *Slope Comparison: Baseline vs Random Slopes Model*



Notes: Each point represents one of 1,484 analyses. The x-axis shows the population slope from the baseline model (common slope across studies); the y-axis shows the population slope from the random slopes model (study-specific slopes). Colors indicate whether dynamics are detected (posterior probability of direction > 0.95) by each model. The dashed line is the 45-degree line.

Table L19 provides additional summary statistics.

Table L19: *Baseline vs Random Slopes Model: Summary*

	Baseline	Random Slopes
Analyses compared	1,484	1,484
Mean slope (SD/month)	-0.0010	-0.0011
Median slope (SD/month)	-0.0021	-0.0018
Mean intercept (SD)	0.277	0.277
Median CI width (slope)	0.1143	0.2371
Dynamics detected (pd $>$ 0.95)	80 (5.4%)	14 (0.9%)
Median $\hat{\tau}_\beta$	—	0.046

Notes: Comparison of population-level slope estimates across 1,484 analyses estimated under both models. The random slopes model adds study-specific slopes $\beta_{ij} \sim N(\beta_j, \tau^2_\beta)$, allowing panel studies to contribute a single noisy slope estimate rather than multiple independent observations.

M Robustness of Meta-Meta-Analysis

This appendix reports robustness of the aggregate fadeout estimate to outlier treatment, effect measure type, and sample restrictions. All specifications use a frequentist mixed-effects model (estimated via REML) with random intercepts and slopes by analysis, crossed study random effects, and precision weights ($1/\text{SE}^2$), approximating the Bayesian model in Section 3.2.

M.1 Outlier Treatment

Table M20 reports estimated fadeout slopes under alternative treatments of extreme effect sizes. The baseline specification uses all observations; subsequent columns progressively winsorize or trim the distribution of Y_{flipped}.

Table M20: *Robustness to Winsorization and Trimming of Effect Sizes*

	Baseline	Win 1/99	Win 5/95	Drop >5 SD	Drop >3 SD	Drop >2 SD
Slope (SD/year)	-0.0527***	-0.0276***	-0.0251***	-0.0292***	-0.0185***	-0.0227***
	(0.0072)	(0.0035)	(0.0029)	(0.0038)	(0.0029)	(0.0029)
Num.Obs.	35097	35097	35097	35013	34836	34427

Notes: Each column reports the coefficient on follow-up time (years) from a precision-weighted mixed-effects regression of Y_{flipped} (effect size in SD units) on follow-up time, with random intercepts and slopes by analysis and crossed study random effects. "Win p/q" winsorizes Y_{flipped} at the pth and qth percentiles. "Drop $> k$ SD" drops observations with $|Y_{\text{flipped}}| > k$. Standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

M.2 By Effect Measure Type

Table M21 estimates the fadeout slope separately by original effect measure type. SMD = standardised mean difference (used directly); MD = mean difference (standardised by dividing by the pooled SD); RR = risk ratio and OR = odds ratio (both converted to SD units via the probit or Chinn transformation described in Appendix E). "Other" pools risk differences, rate ratios, incidence rate ratios, and prevalence ratios.

Table M21: *Fadeout Slope by Effect Measure Type*

	SMD	RR	MD	OR	Other
Slope (SD/year)	-0.0400***	-0.0197***	-0.0418	-0.0410***	-0.0472
	(0.0067)	(0.0047)	(0.0255)	(0.0105)	(0.0582)
Num.Obs.	16481	8596	5777	3863	158

Notes: Each column restricts the sample to observations originally reported in the indicated effect measure and fits the mixed-effects model separately. SMD = standardised mean difference; MD = mean difference; RR = risk ratio; OR = odds ratio. Standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

M.3 Sample Restrictions

Table M22 varies the analysis sample. "Baseline" includes all analyses with any temporal variation in follow-up time. "Moderate" restricts to analyses with IQR of follow-up time $\geq$ 3 months and

$\geq$ 3 studies, matching the analysis-level sample used in Section 3.1. "Panel studies" restricts to observations from studies that have multiple distinct follow-up periods within their analysis (within-study variation in timing). "Has panel" keeps all observations from analyses containing at least one such panel study.

Table M22: *Fadeout Slope Under Alternative Sample Restrictions*

	Baseline	Moderate	Panel studies	Has panel
Slope (SD/year)	-0.0527***	-0.0284***	-0.0712***	-0.0679***
	(0.0072)	(0.0069)	(0.0128)	(0.0116)
Num.Obs.	35097	16485	9379	17282

Notes: Each column varies the estimation sample. "Moderate" requires IQR of follow-up $\geq$ 3 months and $\geq$ 3 studies per analysis. "Panel studies" keeps only observations from studies with multiple follow-up periods within an analysis. "Has panel" keeps all observations from analyses containing at least one panel study. Standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

N Publication Bias

Our fadeout estimates could be contaminated by publication bias if the degree of bias varies with follow-up time. The concern is straightforward: if journals and authors are more likely to selectively report significant results at short follow-up than at long follow-up, then short-run effect sizes will be inflated relative to long-run measurements, and part of what we measure as fadeout would really be the unwinding of selection. To assess whether this is the case, we first ask whether publication bias *signatures* differ across follow-up horizons—using caliper tests for p-value bunching and Egger's funnel asymmetry test within five time buckets—and then directly correct for any bias using the PET-PEESE method. The five buckets are: $t = 0$ (end of treatment; $N = 21,823$), 0–3 months ($N = 6,612$), 3–6 months ($N = 4,935$), 6–12 months ($N = 9,694$), and 12+ months ($N = 5,660$).

Caliper test. The caliper test compares counts of p-values just below and just above conventional significance thresholds; excess mass below indicates selective reporting or outcome switching (Gerber & Malhotra, 2008). Figure N23 plots the caliper ratio for each follow-up bucket. At the $p = 0.05$ threshold with a ±0.01 window, bunching is concentrated at short-to-medium follow-up: the ratio is 1.43 ($p = 0.003$) at 0–3 months and 1.66 ($p < 0.001$) at 3–6 months, but insignificant at $t = 0$ (ratio = 0.95) and at 12+ months (ratio = 0.95). This pattern—stronger bunching at intermediate horizons, weaker at both ends—suggests that selective reporting is most acute for the earliest post-treatment follow-up measurements, where investigators may have the most discretion over which outcomes to highlight.

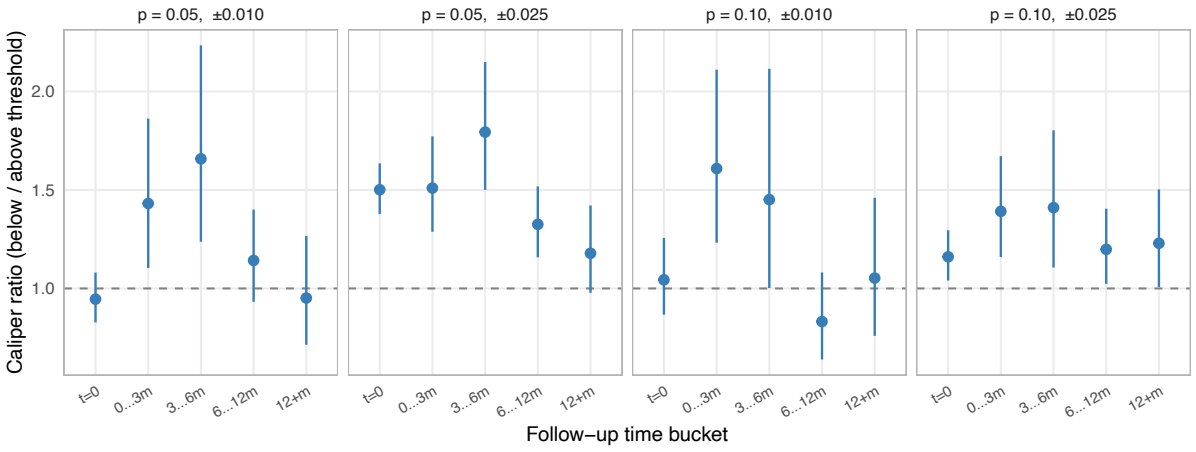


Figure N23: *Caliper Test Ratio by Follow-up Time Bucket*

Notes: Each panel corresponds to a significance threshold and caliper window width. The caliper ratio is the count of p-values just below the threshold divided by the count just above. Ratio > 1 indicates excess mass below the threshold, consistent with selective reporting. Error bars show exact binomial 95% confidence intervals, transformed to the ratio scale. The dashed line marks ratio = 1 (no bunching). P-values computed from $z = Y_{\text{flip}}/\text{SE}$.

Egger's test. Egger's test regresses effect size on its standard error; a positive slope indicates that less precise studies report larger effects, consistent with publication bias (Egger, Smith, Schneider, &

Minder, 1997). Table N24 reports results by follow-up bucket. Within-analysis, the Egger coefficient declines steadily with follow-up time: from 1.00 at $t = 0$ and 1.35 at 0–3 months to 0.58 at 6–12 months and 0.50 at 12+ months (all significant). Funnel asymmetry is roughly twice as severe at short horizons as at long ones, consistent with the caliper test evidence that publication bias disproportionately inflates short-run effect estimates.

Table N24: *Egger's Funnel Asymmetry Test by Follow-up Time Bucket*

Time period	N	(1) WLS	(2) WLS + Analysis FE
$t = 0$	21,823	1.440***	1.000***
		(0.018)	(0.071)
0–3 months	6,612	1.496***	1.353***
		(0.034)	(0.173)
3–6 months	4,935	0.926***	0.970***
		(0.066)	(0.131)
6–12 months	9,694	0.936***	0.578***
		(0.028)	(0.074)
12+ months	5,660	0.908***	0.501***
		(0.042)	(0.135)

Notes: Dependent variable is direction-standardized effect size (Y_{flip}, in SD units). Each row reports the coefficient on SE from $Y_{\text{flip},i} = \alpha + \beta \cdot \text{SE}_i + \epsilon_i$. Column (1): WLS with weights $1/\text{SE}^2$. Column (2): WLS with analysis fixed effects and standard errors clustered at the analysis level (singleton observations dropped). A positive coefficient indicates larger effects in less precise studies, consistent with publication bias. Standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

PET-PEESE correction. Both diagnostic tests point in the same direction: publication bias signatures are stronger at shorter follow-up horizons, exactly the pattern that would inflate apparent fadeout. The question is whether the magnitude is large enough to matter. The PET-PEESE method (Stanley & Doucouliagos, 2014) does both: the Precision-Effect Test (PET) adds SE as a control in the fadeout regression, so that the coefficient on time gives the bias-corrected slope. When PET detects significant bias, the Precision-Effect Estimate with Standard Error (PEESE) refines the correction using SE^2, which better approximates the bias function.

Table N25 reports the results. Within analysis fixed effects, the PET coefficient on SE is 0.8838 ($p = 0$): less precise observations report larger effects, confirming funnel asymmetry. Following the PET-PEESE protocol, we report the PEESE-corrected fadeout slope: −0.001 SD/month (SE = 0.00045), virtually identical to the naive estimate of −0.00101 SD/month. The PEESE SE^2 coefficient is small and insignificant within analyses ($p = 0.166$), indicating that publication bias has negligible impact on the fadeout slope once analysis fixed effects absorb cross-analysis heterogeneity.

Column 8 adds an SE × months interaction to test whether bias differentially affects the slope. The interaction is −0.01193 ($p = 0.037$): less precise observations show steeper apparent fadeout, consistent with significance filtering that inflates initial effects more in underpowered studies. Once the interaction is included, the main effect on time becomes insignificant (0.00026, SE = 0.00072). Adjusting for publication bias does not increase—and, if anything, decreases—the estimated fadeout.

Table N25: *Fadeout Regression: PET-PEESE Publication Bias Correction*

	Naive	PET	PEESE	PET+Int	Naive (FE)	PET (FE)	PEESE (FE)	PET+Int (FE)
Months since treatment end	0.00287***	0.00031***	0.00278***	0.00126***	-0.00101**	-0.00071	-0.00100**	0.00026
	(0.00010)	(0.00009)	(0.00010)	(0.00012)	(0.00045)	(0.00043)	(0.00045)	(0.00072)
SE		1.37266***		1.44034***		0.88376***		0.93867***
		(0.01150)		(0.01247)		(0.07590)		(0.07393)
SE^2			0.32099***				0.05382	
			(0.01190)				(0.03882)	
SE × Months				-0.02132***				-0.01193**
				(0.00153)				(0.00572)
Intercept	0.00101***	0.00010	0.00097***	0.00008				
	(0.00015)	(0.00014)	(0.00015)	(0.00014)				
N	48,803	48,803	48,803	48,803	48,803	48,803	48,803	48,803
R^2	0.016	0.238	0.030	0.241	0.594	0.606	0.594	0.606
Analysis FE	No	No	No	No	Yes	Yes	Yes	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

Notes: Dependent variable is direction-standardized effect size (Y_{flip}, in SD units), precision-weighted ($1/\text{SE}^2$). PET adds SE as a control; PEESE adds SE^2; PET+Int adds SE and SE × Months (Stanley & Doucouliagos, 2014). Columns (5)–(8) include analysis fixed effects with standard errors clustered at the study level. The PET-PEESE protocol reports the PEESE estimate when PET rejects the null of no bias (SE coefficient = 0). $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

O Funding Instrument

This appendix provides methodological details for the funding instrument and control function selection correction presented in Section 4.4.

O.1 Instrument Construction

The instrument exploits variation in NIH disease-area funding across years. The National Institutes of Health tracks annual spending through its Research, Condition, and Disease Categorization (RCDC) system, which classifies spending into 316 categories. We use the 226 "leaf" categories—those at the bottom of the RCDC hierarchy—to avoid double-counting, since higher-level categories aggregate spending from their children. RCDC data cover fiscal years 2008–2024.

To match Cochrane reviews to RCDC categories, we compute text embeddings (OpenAI `text-embedding-3-small`) for both the abstract of each Cochrane review and the description of each RCDC category. For each review i, we compute the cosine similarity sim_{ik} between its embedding and that of each RCDC category k, then convert to weights using a softmax function with temperature $\tau = 0.01$:

$$w_{ik} = \frac{\exp(\text{sim}_{ik}/\tau)}{\sum_{k'} \exp(\text{sim}_{ik'}/\tau)} \tag{42}$$

The low temperature produces near hard-assignment: 62.14% of reviews have a single category receiving more than 80% of the weight, so most reviews are effectively matched to one RCDC category.[7]

The instrument for study i ending in year t is a weighted average of log funding across matched categories:

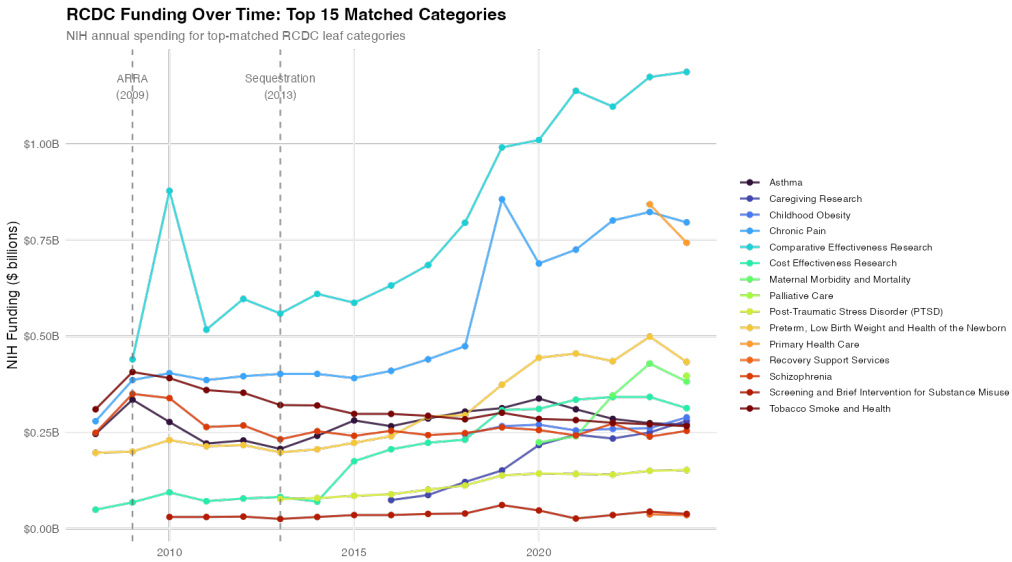
$$IV_{it} = \sum_k w_{ik} \log(\text{funding}_{k,t}) \tag{43}$$

where $\text{funding}_{k,t}$ is NIH RCDC spending on category k in fiscal year t. The instrument is standardized to mean zero and unit variance. The sample is restricted to studies with estimated end dates in 2010–2022, ensuring that contemporaneous year t falls within RCDC coverage. This yields 1,704 studies.

The identifying variation comes from year-to-year fluctuations in NIH categorical spending interacted with the disease-area composition of each Cochrane review. Figure O26 plots RCDC spending over time for the 15 categories most frequently matched to our sample. Key funding shocks include the American Recovery and Reinvestment Act (2009), which provided a temporary NIH budget boost captured for studies ending circa 2011–2012, and sequestration (2013), which imposed an approximately 5% across-the-board cut.

[7] For example, "Family therapy for asthma in children" matches to the RCDC category *Asthma* with weight 1.00; "Back schools for non-specific low-back pain" to *Back Pain* (1.00); and "Training health professionals in smoking cessation" to *Tobacco Smoke and Health* (0.99). Reviews covering broader topics receive more diffuse weights: "User-held personalised information for routine care of people with severe mental illness" splits across *Psychosis* (0.26), *Mental Illness* (0.25), *Cost Effectiveness Research* (0.17), and others.

Figure O26: *NIH RCDC Funding by Category Over Time*



Notes: Annual NIH spending (in billions of dollars) for the 15 RCDC categories most frequently matched to Cochrane reviews in our sample, ordered by mean softmax weight. The ARRA stimulus (2009) and sequestration (2013) are visible as common shocks across categories, but the relative magnitudes vary by disease area.

O.2 Sample

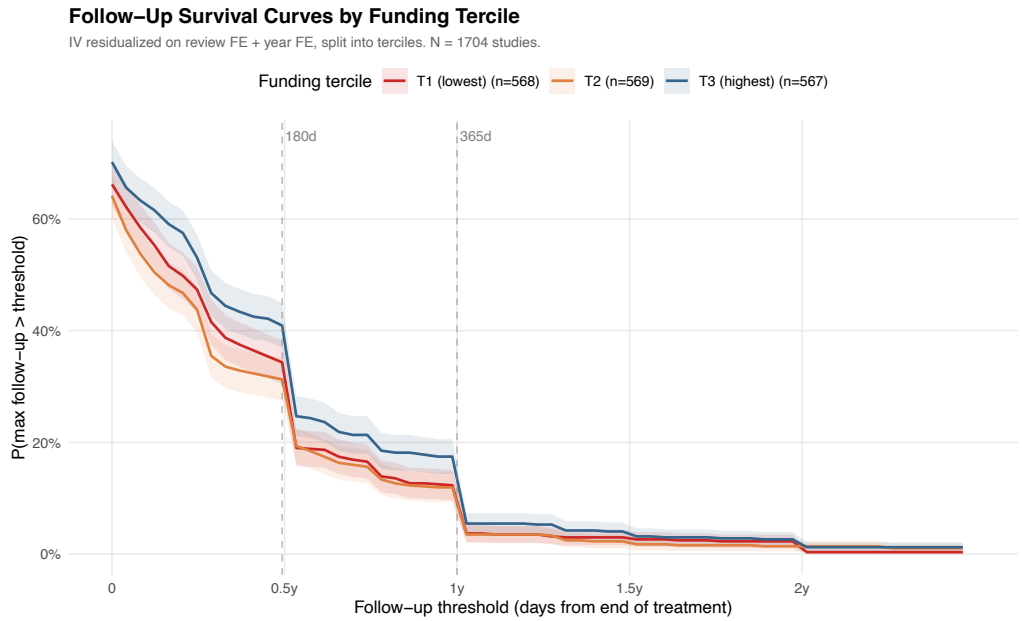
The instrument analysis uses a subsample of the 12,022 unique studies in the psychosocial sample. Calendar end dates are available for 3,517 studies (extracted from Cochrane review characteristics and, where necessary, from the original trial publications). Of these, 2,190 have end dates in the 2010–2022 window where RCDC funding data are available. The sample then narrows further at each stage of the analysis:

Stage	Studies	Notes
End dates in 2010–2022	2,190	RCDC funding window
Valid funding instrument	2,184	6 studies lack RCDC category matches
Non-missing follow-up	1,704	480 studies lack Cochrane timing data
CF second stage	1,472	232 studies have no usable Y/t observations
Observations (linear CF)	8,961	
Observations (DTH CF)	8,944	

The control function subsample covers approximately 17% of the full estimation sample. Mean effect sizes and follow-up times are similar between the CF subsample and the full sample, suggesting that the restriction to studies with RCDC-era end dates does not introduce substantial compositional bias.

Figure O27 provides a non-parametric visualization of the first stage. Studies are split into terciles of the contemporaneous funding instrument, residualized on review and year fixed effects. The top tercile shows visibly higher follow-up survival rates, particularly beyond 180 days.

Figure O27: *Follow-Up Survival Curves by Funding Tercile*



Notes: Each curve shows the fraction of studies with maximum post-treatment follow-up exceeding the threshold on the horizontal axis. Studies are split into terciles of the contemporaneous funding instrument (year t), residualized on review and year fixed effects. Higher funding terciles correspond to studies in reviews that received above-average NIH funding for their disease area in the year the study ended. Shaded bands show 95% confidence intervals.

O.3 Control Function Methodology

O.3.1 Linear Control Function

Error decomposition. To see why selection biases the slope and how the control function corrects it, decompose the outcome error into study-level and idiosyncratic components:

$$Y_{it} = \alpha_0 + \alpha_1 t + \underbrace{\mu_i + \nu_i \cdot t}_{\text{study-level}} + \underbrace{e_{it}}_{\text{idiosyncratic}} \tag{44}$$

where μ_i is a study-specific intercept (some interventions have larger or smaller initial effects), ν_i is a study-specific slope deviation (some interventions fade faster or slower than the population average α_1), and e_{it} is measurement error. The naive OLS estimate of α_1 is biased whenever ν_i is correlated with which values of t we observe for study i. Observations at long horizons come only from studies with long follow-up; if those studies have systematically higher ν_i (less fadeout), then $\text{Cov}(t_{it}, \nu_i) > 0$ and OLS overstates persistence.

Control function. The control function approach (Olsen, 1980) uses the funding instrument to proxy for the unobservable ν_i that drives both follow-up choice and slope heterogeneity. The first stage is a linear probability model (LPM) for the selection indicator $F_i = \mathbf{1}[\text{max follow-up} \geq 365\text{d}]$:

$$F_i = \beta \cdot IV_{it} + \gamma_r + \delta_t + \eta_i \tag{45}$$

where γ_r are review fixed effects and δ_t are year fixed effects.[8] The residual $\hat{\eta}_i = F_i - \hat{F}_i$ captures the component of follow-up timing not explained by funding. Under the exclusion restriction, this residual is correlated with μ_i and ν_i (through the researcher's private information about persistence) but the instrument itself is not.

Assuming that both μ_i and ν_i depend linearly on the selection unobservable η_i, the second stage absorbs the bias by including $\hat{\eta}_i$ and its interaction with time:

$$Y_{it}^{\text{flip}} = \alpha_1 t + \delta_\mu \hat{\eta}_i + \delta_\nu \hat{\eta}_i \cdot t + \gamma_j + \varepsilon_{it} \tag{46}$$

where γ_j are analysis fixed effects and standard errors are clustered by study. The term $\delta_\mu \hat{\eta}_i$ absorbs the correlation between μ_i and η_i (level selection), while $\delta_\nu \hat{\eta}_i \cdot t$ absorbs the correlation between ν_i and η_i (slope selection). After conditioning on both, α_1 is the selection-corrected fadeout rate. If $\delta_\nu > 0$, studies that endogenously chose long follow-up have less fadeout than the population average, confirming that the naive slope understates true fadeout.

O.3.2 Discrete-Time Hazard Control Function

The linear control function treats selection as a binary event: either a study has follow-up beyond 365 days or it does not. But selection operates continuously. A study followed to 6 months contributes observations at short horizons but not at 12 months; a study followed to 3 years contributes everywhere. The selection severity increases with the follow-up horizon, and the 365-day threshold is arbitrary. The discrete-time hazard (DTH) control function addresses this by modeling the full distribution of follow-up duration and constructing a selection correction that varies by horizon.

We divide follow-up into 90-day bins and expand each study into a person-period panel: one row per bin, up to and including the bin in which the study's follow-up ends. For a study followed to 200 days, this produces rows for the first three bins (0–90, 90–180, 180–270), with an "exit" indicator $d_{ik} = 1$ in the third. We then estimate a linear probability model for the hazard of exiting in each bin, conditional on having survived to that bin:

$$d_{ik} = \alpha_k + \beta_k \cdot IV_{it} + \gamma_r + \delta_t + \eta_{ik} \tag{47}$$

where α_k are bin dummies (capturing the baseline hazard shape), γ_r and δ_t are review and year fixed effects, and β_k is a bin-specific coefficient on the funding instrument. Allowing β_k to vary by bin accommodates the possibility that funding affects the exit hazard differently at different horizons — for instance, funding may strongly predict whether a study has any medium-run follow-up but matter less for the decision to continue beyond two years. The LPM residual $\hat{\eta}_{ik} = d_{ik} - \hat{h}_{ik}$ captures whether study i exited follow-up in bin k earlier or later than its funding environment predicts.

Each observation Y_{it} is mapped to the bin containing its follow-up time t and assigned the corresponding residual $\hat{\eta}_{ik}$. Unlike the linear control function, where $\hat{\eta}_i$ is a single study-level value,

[8] We use an LPM rather than a probit because many reviews contain no studies with $F_i = 1$, which would cause perfect separation in a probit. The LPM handles these constant-outcome groups naturally.

$\hat{\eta}_{ik}$ varies across observations within a study: at each horizon, it reflects the study's deviation from predicted follow-up behavior at that specific time point. The second stage replaces $\hat{\eta}_i$ with $\hat{\eta}_{ik}$ and $\hat{\eta}_i \cdot t$ with $\hat{\eta}_{ik} \cdot t$.

Standard errors. Standard errors in the second stage are cluster-robust at the study level but do not account for estimation uncertainty in the generated regressors ($\hat{\eta}_i$ or $\hat{\eta}_{ik}$). A bootstrap procedure resampling studies and re-estimating both stages is standard for correcting this; the reported standard errors should be interpreted as lower bounds on the true uncertainty.

Table O28: *First Stage by Country: US Interaction*

	P(≥90d)		P(≥180d)		P(≥270d)		P(≥365d)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Funding t	0.069	-0.015	0.067	0.074	0.095	0.219*	0.128**	0.245**
	(0.069)	(0.121)	(0.066)	(0.123)	(0.062)	(0.123)	(0.059)	(0.120)
US study	-0.034	-0.030	-0.025	-0.031	-0.042	-0.046	-0.043*	-0.047*
	(0.033)	(0.038)	(0.031)	(0.036)	(0.026)	(0.031)	(0.023)	(0.028)
Funding t × US	-0.001	-0.002	-0.013	-0.011	-0.004	-0.006	0.006	-0.000
	(0.028)	(0.034)	(0.027)	(0.032)	(0.025)	(0.030)	(0.023)	(0.028)
N	1,704	1,704	1,704	1,704	1,704	1,704	1,704	1,704
R^2	0.414	0.501	0.435	0.517	0.313	0.399	0.327	0.405
Mean dep. var.	0.483	0.483	0.356	0.356	0.178	0.178	0.134	0.134
Review + Year FEs	✓	✓	✓	✓	✓	✓	✓	✓
Review × trend		✓		✓		✓		✓

Notes: The unit of observation is a study. The dependent variable is an indicator for whether the study's maximum post-treatment follow-up exceeds the stated threshold. Each column reports a linear probability model with the contemporaneous funding instrument (year t), a US study indicator, and their interaction. The interaction tests whether the instrument has differential predictive power for US studies, where NIH is a primary funder. Odd columns include review and year fixed effects; even columns additionally include review-specific linear time trends. HC1 standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.