Vitamin A Supplements and Child Mortality: Resolving A Controversy In Meta-Analysis

Rafe "Rachael" Meager[*] Witold Więcek[†‡]

October 27, 2025

Abstract

Vitamin A supplementation is considered one of the most effective interventions to reduce child mortality in developing nations. However, its estimated effect varies substantially across studies, making the results of meta-analyses in this literature highly sensitive to methodological choices.

We compare the theoretical properties and empirical performance of classical fixed-effects and random-effects meta-analysis models on simulated bodies of evidence. We consider Bayesian implementation using a hierarchical model. We find that using random effects either matches or substantially outperforms the fixed effects method in terms of mean squared error and is robust to misspecification of the likelihood.

Applied to a set of 18 studies, a Bayesian model estimates that vitamin A supplementation reduces mortality risk by 25% (95% interval 40% to 11%), compared to 12% under the fixed-effects model (95% interval 7% to 17%).

The model suggests that the underlying heterogeneity across studies is large (66% of the cross-study variation attributable to genuine differences in treatment effects), but this is not precisely estimated (95% interval 26% to 91%). Fixed effects thus underestimate both the reduction in mortality and the uncertainty surrounding the general effect of supplementation.

1 Introduction

Public health researchers and international aid organisations consider vitamin A supplementation one of the most important and cost-effective interventions to reduce child mortality (Imdad et al 2022, GiveWell 2024). UNICEF refers to vitamin A programming

[*]University of New South Wales. Contact: r.meager@unsw.edu.au.

[†]Development Innovation Lab, University of Chicago

[‡]We thank Abhijit Banerjee, Esther Duflo, Anna Mikusheva, Victor Chernozhukov and Alfred Sommer for helpful discussions and suggestions. We also thank Joshua Muthu for research assistance. All remaining mistakes are our own. This is a perpetual working paper. Please send critiques and corrections via email.

as a "prerequisite" for achieving Millennium Development Goal 4, the goal of child survival, particularly in countries with high vitamin A deficiency (VAD) rates (UNICEF, 2007). The biological mechanism by which vitamin A affects mortality is relatively well understood: retinoic acid has a specific role in visual system function and in the immune system where it improves T- and B-cell gut-homing capacity and enhances T-cell proliferation (Mora et al, 2008 and Iwata et al 2004). This explains why vitamin A deficiency is often associated with greater mortality risk in longitudinal data (Sommer et al, 1983).

The potential of large-scale vitamin A supplementation to effectively reduce child mortality was first established in Sommer et al (1986) using a randomized controlled trial (RCT) in Sumatra, Indonesia; the program reduced child mortality by 34% of base risk. The literature that followed displayed substantial variation in results across different contexts, on balance showing positive effects (e.g. Vijayaraghavan et al 1990, Rahmathullah et al 1990, Herrera et al 1992, Arthur et al 1992, VAST 1993). While the effect of supplementation in the absence of vitamin A deficiency is unclear, this deficiency is prevalent in so many developing countries that the World Health Organisation recommends supplementation among children 6-59 months in settings where there is VAD (WHO, 2016).

However, results from the DEVTA trial, the largest ever randomised controlled trial of vitamin A supplementation, threatened to overturn this consensus (Awasthi et al, 2013). The study, performed in India, reported only a small and statistically insignificant reduction in child mortality. This was not the first time a study reported no effect of vitamin A, and one previous study even found a slight and insignificant increase in mortality (Herrera et al, 1992). But Awasthi et al (2013) studied one million subjects and produced the most precise estimate of the risk ratio ever recorded. However, despite this impressive precision, the DEVTA trial was different from previous trials in many ways: in particular, it had a different delivery mechanism and a lower rate of compliance (Awasthi et al 2013, Sommer et al 2013). As a result, some researchers remained unconvinced that the DEVTA trial should meaningfully overturn or alter the conclusions of the previous three decades of research on Vitamin A (Sommer et al, 2013, and Garner et al, 2013).

Faced with disagreements in the scientific community about the implications of a body of research, the methodological question of how to aggregate and interpret the evidence from seemingly-contradictory studies becomes crucially important. This question has material consequences for science and policy, because the conclusions of evidence aggregation exercises are often sensitive to the statistical methodologies chosen in these contexts. The most recent Cochrane review by Imdad et al (2022) uses a fixed-effects meta-analysis model and finds reduction in risk of mortality of 12% with narrow confidence intervals (RR = 0.88, 95% CI 0.83 to 0.93). This is similar in magnitude to the meta-analytic result by Awasthi et al (2013) of 11%, which combined DEVTA with prior trials, but a difference of more than ten percentage points from the meta-analyses using random-effects (reported as sensitivity analyses by Mayo-Wilson et al 2011, Imdad et al 2022).[1]

Even though both models directionally agree and yield statistically significant results,

[1]In the economics literature, terms "constant effects" and "random coefficients" are used for fixed and random effects respectively.

that is still a two-fold difference in magnitude, which is presumably of major importance to public health decision makers. Which model is appropriate for this setting? While Awasthi et al (2013) claim that random effects meta-analytic techniques overweigh small studies, and "conceal the reliability" of large studies such as DEVTA, Sommer et al (2013) argue that given the heterogeneity in the results a random-effects aggregation model is more appropriate. *A priori* there are clear reasons to expect heterogeneity in effects; for example, some trials use vitamin A single megadose, while others use weekly low-doses; baseline levels of VAD and compliance levels also vary widely.

However, because RE assigns higher weights to small studies, its results can clearly be biased in presence of some small-study effects. In other words, while smaller studies may systematically have larger effects due to real heterogeneity (e.g. plausible mechanisms as to why smaller RCTs conducted in areas where potential treatment effect is higher), which makes RE model a more appropriate choice, small-study effects may also arise due to publication bias towards positive or significant results in VAS literature. Inspection of funnel plots in two previous meta-analyses by Mayo-Wilson et al and Imdad et al does suggest asymmetry, which can be indicative of this problem. While both meta-analyses conclude (in light of funnel plot inspection and risk of bias assessments) that the positive results on mortality reductions are not nullified by that finding, any meta-analytic model should also consider adjustments for potential publication bias.

In this paper we aim to contribute to the VAS debate by investigating which method should be used to aggregate the evidence in the vitamin A literature, and apply that method to perform a new meta-analysis.

First, we review the theoretical properties of fixed and random effects estimators and assess their performance on simulated literatures. Earlier work by Hedges and Vevea (1998), Overton (1998), and Field (2001) reported simulation studies assuming that the functional form of either the fixed-effects (FE) model or the random-effects (RE) model was exactly correct, and scored performance by type 1 error rate. By contrast, we examine performance on a wider range of data-generating processes for which neither meta-analytic model is exactly correct, including distributions with fat tails and multi-modal mixture distributions. To assess performance, we conduct a Monte Carlo simulation and compare the mean squared error (MSE) of both procedures in estimating the average treatment effect. This approach is chosen over evaluating hypothesis test error rates, because research communities are increasingly interested in getting an estimate as close as possible to the true effect size, not simply testing for the existence of a non-zero effect (Wasserstein and Lazar 2016).

We then fit both FE and RE models using data from the most recent meta-analysis by Imdad et al. (2022) and conduct a formal comparison of these two models. Lastly, we examine publication bias and apply adjustments to the overall effect sizes that take it into account.

2 Methodologies for Meta-Analysis

2.1 Fixed Effects

The canonical fixed-effects meta-analysis model uses an inverse-variance weighted average of the studies' estimated effects to produce an estimate of a general or typical treatment effect (Higgins and Green, 2011, 9.4.3). For a set of K studies with estimated treatment effects and standard errors $\{\hat{\tau}_k, se_k\}_{k=1}^K$,[2] the FE estimator is

$$\hat{\tau}_{FE} = \sum_{k=1}^{K} \hat{\tau}_k \frac{(se_k^2)^{-1}}{\sum_{k=1}^{K}(se_k^2)^{-1}} \quad . \tag{1}$$

The standard error of this estimate is calculated as follows:

$$\hat{se}_{FE} = \sqrt{\frac{1}{\sum_{k=1}^{K}(se_k^2)^{-1}}} \quad . \tag{2}$$

If the studies are estimating the same underlying treatment effect τ, and each study's estimated effect $\hat{\tau}_k$ is asymptotically normally distributed, then this method is optimal in the class of estimators produced by linear combinations of the $\{\hat{\tau}_k\}_{k=1}^K$. This result is proved in Hedges and Vevea (1998) but had been established informally in the statistics literature years prior. We review it here: formally, assume that the point estimators are consistent and that each study has enough data and the empirical distributions have enough regularity such that, for all k, $\hat{\tau}_k \sim \mathcal{N}(\tau, se_k^2)$.[3]

Now consider the class of estimators for which $\hat{\tau} = \sum_{k=1}^{K} w_k \hat{\tau}_k$ for for some positive weights, $\sum\{w_k\}_{k=1}^K = 1$. The sampling distribution of any such estimator will be

$$\hat{\tau} \sim \mathcal{N}(\tau, \sum_{k=1}^{K} w_k^2 se_k^2). \tag{3}$$

Since all such estimators are consistent by construction, we seek estimator with the minimal mean squared error (MSE). Taking a first order condition with respect to the weights shows that

$$\left\{ \frac{(se_k^2)^{-1}}{\sum_{k=1}^{K}(se_k^2)^{-1}} \right\}_{k=1}^{K} = \underset{\{w_k\}_{k=1}^K}{\arg\min} \sum_{k=1}^{K} w_k^2 se_k^2. \tag{4}$$

This result demonstrates the efficiency of the FE estimator in this class. Notice that the standard errors were ancillary statistics for τ, yet are necessary to construct the efficient estimator.

[2] Some of the statements we will make about optimality of estimators will require for within-study variances to be correctly specified; even though in practice we use estimators of se_k reported by the studies, their variance is low from a practical viewpoint the standard errors can be treated as known.

[3] An assumption of this sort is typically made during the calculation of most standard parametric confidence intervals and p-values in the medical and social science literature.

The FE model can be implemented using either a frequentist or a Bayesian approach. While the former is typically used, the Bayesian framework allows for direct comparison between fixed- and random-effects models, which we will discuss below.

2.2 Random Effects Methods

If each $\hat{\tau}_k$ estimates a different τ_k rather than one common τ, the FE estimator is no longer optimal. To see this, consider a hypothetical case in which the $\{\hat{\tau}_k\}_{k=1}^K$ are estimates of K distinct parameters according to the following hierarchical likelihood:

$$\begin{aligned} \hat{\tau}_k &\sim \mathcal{N}(\tau_k, se_k^2) \ \ \forall\ k \\ \tau_k &\sim \mathcal{N}(\tau, \sigma_\tau^2) \ \ \forall\ k. \end{aligned} \tag{5}$$

In this model, each $\hat{\tau}_k$ actually measures a k-specific underlying effect τ_k, but these are centered around a common τ component. As before, we seek the minimum variance estimator of this τ given the estimates $\{\hat{\tau}_k, se_k\}_{k=1}^K$, in the class of estimators taking the form $\hat{\tau} = \sum_{k=1}^K w_k \hat{\tau}_k$ for some positive weights, $\sum\{w_k\}_{k=1}^K = 1$.

By law of total variance and the properties of sum of normal distributions,

$$\hat{\tau} \sim \mathcal{N}(\tau, \sum_{k=1}^K w_k^2 (se_k^2 + \sigma_\tau^2)). \tag{6}$$

Since the estimator is consistent, the combination of weights that minimises MSE will lead to an efficient estimator. By the same argument that produced the optimal weights in the FE model,

$$\left\{ \frac{(se_k^2 + \sigma_\tau^2)^{-1}}{\sum_{k=1}^K (se_k^2 + \sigma_\tau^2)^{-1}} \right\}_{k=1}^K = \underset{\{w_k\}_{k=1}^K}{\arg\min} \sum_{k=1}^K w_k^2 (se_k^2 + \sigma_\tau^2). \tag{7}$$

Using these weights implements a random effects (RE) estimation method. When $\sigma_\tau^2 \neq 0$, the FE estimator diverges from this estimator and has higher variance: the FE method is thus inefficient in this case. It is only optimal when the data from all K samples or studies estimates exactly the same effect. Conversely, when the true σ_τ^2 is very large, the FE estimate may be very misleading.

In contrast, the RE estimator in (7) accounts for underlying heterogeneity and thus tempers the influence that any single study can have on the estimate of τ. The RE estimator is optimal if σ_τ^2 is known and if the effects are independently, identically, and normally distributed around their common mean. In practice, even if this functional form is correctly specified, σ_τ^2 is not known and must be estimated. As described by Rubin (1981), it can be estimated via Bayesian inference or via Maximum Likelihood (which is often called "empirical Bayes" in this case). For relatively simple models, such as (5), the performance, computational complexity, and running time are similar for the two methods.

In more complicated models, the ML estimation quickly becomes a challenging multivariate optimisation problem that requires inversion of large matrices to calculate standard errors. In these more complex cases, the "Empirical Bayes" MLE is performed using a 2-step procedure that does not accurately quantify the uncertainty in the unknowns. In general, the commonly used frequentist random-effects meta-analytic estimators are known to underestimate heterogeneity.[4] In some cases, these estimators "snap to boundary" (i.e. estimate the heterogeneity parameter τ as zero) even when heterogeneity is present, but highly uncertain. This occurs particularly often when the number of studies in a meta-analysis is low.

Bayesian hierarchical model

Our preferred alternative is to conduct inference on model (5) using a Bayesian hierarchical model (BHM).

As with all Bayesian models, using this type of inference improves model checking and allows for generating posterior predictive quantities. BHMs also avoid the boundary estimate problem faced by frequentist estimators, and can improve overall inference on heterogeneity by introducing informative or mildly informative priors (Chung et al, 2012). [5]

In this paper we estimate BHM using Hamiltonian Monte Carlo with a No-U-Turn-Sampler using Stan software.[6] This Bayesian approach to meta-analysis is implemented in an accessible way in the R package *baggr*, which in addition to model (5) also includes other commonly used meta-analysis models (Więcek and Meager, 2022). The package also includes fitting of FE models, calculation of pooling statistics, and model selection through cross-validation, which we describe below.

Pooling metrics

To formally quantify the heterogeneity in BHM we can compute the conventional pooling factor, as specified in Gelman et al (2006):

$$\omega(\tau_k) = \frac{\hat{se}_k^2}{\hat{se}_k^2 + \hat{\sigma}_\tau^2} \tag{8}$$

When values of the pooling factor are close to 1, sampling variation in studies is much larger than cross-site heterogeneity. When pooling is low, cross-site heterogeneity

[4] This is true of most but not all estimators; some are known to consistently underestimate heterogeneity. These problems occur especially when sample sizes are small.

[5] These priors can be derived from other meta-analyses in similar domains of study. See, for example, work by Turner et al, 2015 on deriving such priors for medical interventions based on Cochrane database of systematic reviews.

[6] Quality of approximation of the posterior distribution by the Monte Carlo methods depends not only on algorithm but also parameters such as the number of iterations (samples from the posterior). While in some cases this requires careful specification and troubleshooting, for meta-analysis models such as in this paper obtaining good convergence is typically easy to achieve. In this paper we use the standard Stan settings and check convergence via the $\hat{R}$ statistic proposed by Gelman and Rubin (1992).

dominates. This factor is the formal complement of the often-used I^2 metric (i.e. $I^2 = 1 - \omega(\tau_k)$), reviewed favourably by Higgins and Thompson (2002), and provides the same information.

Posterior predictive distributions

The Bayesian approach to meta-analysis also allows for a straightforward calculation of the effect's posterior predictive distribution (p.p.d.). Generically, the p.p.d. tells the researcher what values the future data may take, conditional on observed data. In the case of meta-analyses, where we are interested in true treatment effects, this refers to a prediction of the effect in "one additional study" or "the next study" , τ', and therefore the generalizability of the intervention. Under the FE model, the p.p.d. is trivially equal to the distribution of the mean treatment effect (since they are shared by each study). In other words, all of uncertainty in the p.p.d. is uncertainty in the estimate of the common effect. Under the RE model, variance is a sum of this uncertainty and the genuine heterogeneity across settings (σ_τ^2).

Cross validation approach to model selection

Using p.p.d's calculated under Bayesian implementation of FE and RE models offers researchers a possibility to compare the two models. We follow Gelman et al. (2014) in implementing leave-one-out cross-validation, a method of estimating model's out-of-sample performance. This is done by excluding one study at the time, refitting the model to remaining studies, and calculating predictive performance in terms of expected log predictive density (elpd).

Difficulty in choosing between RE and FE models

Even a researcher who is confident of heterogeneous effects cannot be certain of the distribution around the common effect τ. The possibility of model error in the specification of the hierarchical model introduces some concern about the actual efficiency of any chosen RE estimator. This problem does not imply that we should use the FE model instead, since if the underlying effects are heterogeneous then the FE estimator is also inefficient. Nor can we typically rely on popular tests of homogeneity in effects, as they often have insufficient power to detect heterogeneity (Field, 2001).

In practice, particularly for experiments performed in the field across different countries, researchers are often unwilling to rule out heterogeneity completely yet unhappy to draw inferences that are based on a potentially misleading specification of that heterogeneity. The relevant question then is which of the two meta-analytic approaches is likely to perform better in the presence of heterogeneity under a variety of conditions and, in particular, when the RE model misspecifies the distribution of treatment effects across studies.

2.3 Publication bias

Publication bias is likely based on what was reported by Imdad et al (2022) and Mayo-Wilson et al (2011) meta-analyses. The weight of evidence that we give to small studies is closely related to the issue of publication bias towards positive or significant results, which may lead to overrepresentation of small studies in our sample. Of course, even without these concerns the publication bias will bias the treatment effect estimates. Therefore we have two aims. First, we assess if the publication bias is present using qualitative and quantiative methods. Second, we calculate an "adjusted" treatment effect estimate: one obtained under a model which allows for publication bias.

2.3.1 Assessment of small-study and publication bias

We first use graphical and formal diagnostics that are standard in meta-analysis. Funnel plots display study-level effect estimates against their standard errors. Visually apparent asymmetry can arise from selective non-reporting of small, non-significant studies, as well as from genuine relationship between effect size and study size, as well as from large unexplained heterogeneity in treatment effects. Noting asymmetry in funnel plots does not allow us to conclude which of these mechanisms may be present, but it may be indicative of the problems we listed.

We use two established tests for funnel-plot asymmetry by Egger (1997) and Begg-Mazdumar. The Egger regression test asks whether, across studies, effects systematically vary with their precision; a non-zero intercept is evidence that smaller, less precise studies tend to report different effects than larger, more precise studies. This pattern is consistent with both small-study effects and selective reporting, and it matters because it implies the pooled estimate may be pulled away from what the largest, most informative trials indicate. The Begg–Mazumdar test complements this by checking whether there is a monotone association between study effects and their variances using rank correlation; a significant correlation likewise signals asymmetry of the funnel and potential small-study effects. Both tests are interpreted cautiously when heterogeneity is substantial or when the number of studies is small; we report them alongside the funnel plots.

2.3.2 Adjustment for publication bias

Where publication bias is plausible, it is often modeled as selection on statistical significance. In the simplest version, we parameterize selection with $\omega \in (0, 1]$, the *relative* publication probability for studies whose test statistic is non-significant (conventionally $|z| \equiv |\hat{\tau}_k|/se_k \leq 1.96$, which corresponds to 5% significance level) compared with significant studies ($|z| > 1.96$). Thus $\omega < 1$ indicates that non-significant results are less likely to be observed. This is a selection (weight-function) model in the spirit of Hedges (1992), and we implement it using the frequentist identification strategy of Andrews and Kasy (2019). We assume a symmetric, two-sided threshold of 1.96 that does not favor the sign

Under the selection model with random effects, we jointly estimate ω along with τ' and σ'_τ, the *publication-bias–adjusted* hypermean and between-study standard deviation.

Intuitively, the observed set of studies is a selectively sampled subset of all conducted studies; the observed-data likelihood is therefore proportional to the standard random-effects likelihood reweighted by the selection probabilities, with a normalizing constant that depends on ω. Following Andrews and Kasy maximum likelihood approach we obtain and report $(\hat{\tau}', \widehat{\sigma_\tau'}, \hat{\omega})$. If $\omega < 1$, this adjustment will typically shrink the pooled effect toward zero and may reduce the estimated heterogeneity; however, the direction and magnitude are empirical and not guaranteed in finite samples. We compare the adjusted quantities to their unadjusted (Bayesian random-effects) counterparts.

2.4 Monte Carlo Simulations

Statistical models rarely describe the true generating process of any data set. For most meta-analytic problems, neither the RE nor the FE estimator will be exactly optimal. But one procedure could still outperform the other under certain conditions, and their relative performance in various settings should inform the choice of estimator. In this section we use a Monte Carlo method to examine the relative performance of RE and FE models on a variety of data-generating processes (DGPs).

We begin by generating data from the RE model with Gaussian distribution of true effects, at different values of standard errors and cross-study variations. We then gradually move away from this model by examining Student-t distributions of effects, then adding location outliers, and finally adding outliers in the standard error distribution as well. Further details of the DGPs are in Appendix B. For each DGP we run a total of $S = 5000$ simulations. We set $K = 18$ to match the size of the vitamin A meta-analysis in Imdad et al (2022).[7]

In each simulation, indexed by s, we collect $\{\hat{\tau}_k^s, se_k^s\}_{k=1}^1 8$, used for fitting models, and set aside true values of effect and heterogeneity, (τ^s, σ^s). We calculate the FE estimator using (1) and (2). For the RE estimation, we fit a BHM model implementing the likelihood in (5) as described above, with diffuse priors.[8] We use a Gaussian prior on the mean and a uniform prior on the heterogeneity:

$$\begin{aligned} \tau^s &\sim \mathcal{N}(0, 10000^2) \\ \sigma_\tau^s &\sim U(0, 1000) \end{aligned} \tag{9}$$

The MSE for FE and RE (BHM) models is then calculated as the mean over S simulations. We summarize the results by looking at ratios of MSEs in FE and RE model. The full results are shown in the Appendix and summarized in graphs that follow. First, consider the normal-normal model as data-generating process (DGP), in which both

[7] Repeating the simulations using $K = 9$, which is the number of trials in Awasthi et al (2011), leads to similar results.

[8] We make no adjustment to the BHM in any of the simulations. Indeed, to make the point that Bayesian hierarchical models do not require good prior information to perform well, the priors are all incorrect in our simulations - which is to say, they are not the distributions from which we draw the parameters in the simulation. However, these priors are so diffuse that this should not meaningfully impede the Bayesian model's performance (and nor do they).

the estimators and true effects are normally distributed. In this case the BHM has the correct functional form. Figure **??** shows the ratio of the FE MSE to BHM MSE for 16 simulations, varying the average standard errors of the point estimates and the average standard deviation of the true effects (σ_τ). Darker red values indicate a higher ratio, which means the FE MSE is much higher than the BHM MSE, an indicator of poor relative performance of the FE model.

We find that RE model performs better than FE for the normal-normal DGP (Figure **??**, top-left panel). A reasonable complaint about the exercise above is that the BHM's superior performance was potentially due to the correct functional form specification. Therefore we now consider deviations from the normal-normal DGP. Overall the BHM still out-performs the FE estimator on average despite the functional form misspecification.

The two methods perform similarly well when the point estimates have standard errors in the range [5,20]: the ratio of MSEs lies between 0.99 and 1.1 in this area, representing at most a 1% improvement from using the FE estimator which occurs when the standard deviation of effects σ_τ lies in [0,5]. When σ_τ lies in [5,20], we can achieve 10% improvement in MSE from using the BHM. However, when the standard errors lie in the range [0,5] the BHM is twice as efficient when σ_τ is also in this range, and up to four times as efficient when σ_τ is in [15,20]. This means that in these squares of the grid, the FE estimator has an increase in MSE of 300% relative to the BHM. By contrast, in any given square of the grid the BHM has at most a 1% increase in MSE relative to the FE when it is beaten by the FE method, which only occurs when σ_τ is very low.

It may seem surprising that the BHM obtains a much lower MSE than the FE Estimator in the case where σ_τ lies in [0,5], which includes the case of no underlying heterogeneity. However, recall that the BHM nests the case of no heterogeneity: the model is perfectly capable of estimating σ_τ^2 to be essentially zero and then effectively implementing the FE model. The BHM uses up a degree of freedom to perform this estimation, but its overall performance is still close to that of the FE estimator in this case. Of course, as shown in table **??** in the Appendix, when $\sigma_\tau = 0$ the FE estimator does have a lower MSE than the BHM estimator: it must do, because correctly setting $\sigma_\tau^2 = 0$ ex-ante is more efficient than estimating it to be 0. However the improvement is small in most of these simulations, and the largest relative improvement (a reduction by half) occurs when there is very little sampling error at all, so in practice this translates to very small absolute improvements. By contrast, the BHM can achieve major reductions in MSE when σ_τ is in the range $(0, 5]$, so overall it has less than half the MSE of the FE on average in the range $[0, 5]$.

A reasonable complaint about the exercise above is that the BHM's superior performance was potentially due to the correct functional form specification. Therefore we now consider deviations from the normal-normal DGP. First, we examine the case where the true effects are Student t distributed across sites with considerably heavy tails, although the sampling uncertainty is still Gaussian. Figure **??**.2 shows the results for a student t with 3 degrees of freedom, which has extremely fat tails relative to the normal (in fact the

kurtosis is infinite).[9] The FE estimator now performs slightly less poorly in the area where the standard errors are in the range [0,5], although it still has double or triple the MSE of the BHM in this area. When the standard errors are in the range [5,20] the FE estimator can now attain up to a 10% reduction in MSE relative to the BHM in some areas of the grid, but the BHM still attains a 10% reduction in other areas. Overall the BHM still out-performs the FE estimator on average despite the functional form misspecification.

We now consider an alternative specification error in which the underlying data-generating process is a mixture distribution with potentially multiple modes. We first use a mixture distribution to create a classical outlier problem, which occurs when one site's true effect takes a value very different from the rest of the sites' true effects. However, in meta-analysis there is another potential type of outlier: one site may have a standard error that is very much smaller or larger than the rest of the sites' standard errors. We refer to this second type of outlier using the terminology "precision outliers", to distinguish from the more classical "location outliers". To my knowledge this paper constitutes the first attempt to formalize this problem, although meta-analytic issues created by differential precision have been noted in Higgins and Green (2011). We examine both types of outliers separately in the following simulations.

Figure **??**.3 shows the ratio of FE MSE to BHM MSE when the normal-normal DGP model has one site in which the true effect is generated by a normal that has been shifted away from the parent mean τ. The results look quite similar to the normal-normal DGP results, with only slightly worse performance from the FE estimator in the upper left quadrant.

Figure **??**.4 shows the ratio of FE MSE to BHM MSE when the normal-normal DGP model has one site in which the standard error is ten times smaller than the rest of the sites' standard errors on average. This makes a substantial difference in the relative performance of the FE estimator - it has double or triple the MSE of the BHM estimator in most of the grid area. In the worst case scenario for the FE it has MSE almost 6 times the size of the BHM estimator's MSE. In the worst case scenario for the BHM estimator, it still has roughly the same MSE as the FE estimator. Thus, in cases where one site is much more precise than all the others and there is no reason to believe effects are homogeneous, the BHM strongly outperforms the FE estimator.

Why does the fixed effects method perform so poorly in this case? While precision is unequivocally a good thing in a single study, it can be problematic in meta-analysis because precision and generalizability are two distinct but related properties. An estimate can be precise - and thus contain a lot of information about the site it comes from - without containing *generalizable* information. In these simulations, the size of the effect which was much more precise was the same as the other effect sizes on average. All the site effects were a priori equally generalizable, and ex-post equally likely to be good indicators of the general effect τ. But the highly precise effect was treated as more generalizable by

[9]The infinite kurtosis makes the simulations unstable. We tested several seeds and did not find major qualitative differences in the ratio of MSEs, although the raw MSEs did take different values even with large numbers of simulations.

the FE estimation process because this method has no way to distinguish precision and generalizability: indeed, if all effects are the same then precision actually is a perfect indicator of generalizability.

By contrast, random effects methods can and must distinguish these two concepts because they must handle the possibility of heterogeneous effects. The BHM uses the concept of a parent distribution, and specifically the functional form of the parent distribution, to place each effect in context of the other effects. In the real world, the precision outlier is likely to be even more misleading for the FE estimator because that site's effect size is probably qualitatively different to the other sites. A study that is ten times more precise than all other studies may have a larger sample size because it is easier to collect data there, or it is more attractive to governments or NGOs to fund big studies there - these factors are probably correlated to the study's true effect size in some way. Thus, the relative reduction in MSE gained by using RE models in the presence of precision outliers is likely to be even greater than the results of these simulations suggest.

3 Vitamin A Supplements Reduce Child Mortality

3.1 Data

In this section we repeat the meta-analysis of the impact of vitamin A supplementation on child mortality by using the same set of papers as the most recent study by Imdad et al (2022). Data are shown in Table **??**. As it is not easy to interpret a log standard error, we also display the upper and lower bounds of the confidence intervals implied by this log(se).

In health literature, it is typical to model treatment effect on events data by using odds ratios (OR), risk ratios (RR), incidence risk ratios (IRR) or hazard ratios (HR). For example, in the vitamin A literature, including the meta-analysis models we cited, risk ratios between treatment and control groups are modeled. Since RR is approximately normal on logarithmic scale, we can apply the BHM from Equation (5). Data are shown in Table **??**. As it is not easy to interpret a log standard error, we also display the upper and lower bounds of the confidence intervals implied by this log(se).

Before proceeding, we note that an earlier meta-analysis by Mayo-Wilson et al. (2011) contains a very similar set of studies. Awasthi et al (2013) include only nine papers. We also repeated our analysis using the inputs from that paper to allow for direct comparison. The results for both RE and FE models are similar to Imdad et al. (2022) and we reach the same conclusions, therefore we do not show them here.

3.2 Fixed-effects and random-effects models

The previous theoretical analysis and simulation studies suggest that a random effects analysis using a Bayesian hierarchical model is typically a more appropriate choice of methodology for aggregating evidence relative to the FE model. While the FE estimator

can beat the BHM when heterogeneity is small relative to sampling error, it rarely achieves a major reduction in MSE in the simulation study presented here. By contrast when the BHM beats the FE model it routinely achieves an MSE half as large, and in some cases can achieve an MSE less than 1/5th the size of its competitor's MSE. This is a very favourable efficiency versus robustness tradeoff, giving a strong argument to consider the RE model implemented via BHM as default choice for meta-analysis.

The vitamin A literature is particularly suited to aggregation using RE because there is scientific and statistical evidence that underlying effects are heterogeneous across contexts and that this heterogeneity is likely large. Vitamin A studies we consider here differ in important ways, spanning several different countries, time periods, implementation and distribution methods and base rates of vitamin A deficiency in the populations studied. Supplementation only has theoretically-grounded medical benefits for individuals with a vitamin A deficiency. As a result, the rate of deficiency in the population is likely to have a major impact on the observed effect of supplementation on the population's risk ratio.

There is also statistical evidence for heterogeneity, primarily that the 95% confidence intervals for the risk ratios do not overlap in many of the studies. Both the DEVTA (2013) study's interval and the Herrera (1992) interval do not overlap with both the Rahmathullah (1990) interval and the Arthur (1992) interval. Indeed, the χ^2 test results presented in Webtable 3 of Awasthi et al (2013) show strong evidence of heterogeneity: the authors do not discuss these results, but the low p-values in this context are indicative of heterogeneity (Cochrane Handbook Section 9.5.2). Moreover, the DEVTA study (Awasthi et al 2013) is almost an order of magnitude more precise than the other studies in the literature. This study is a precision outlier, and as we saw in the simulations, it is preferable to use an RE model in such a case.

All of the above suggests that a random-effects model is an appropriate choice. We therefore fit both FE and RE models using Bayesian inference, which in turns allows us to compare them in terms of expected log predictive density. We use weakly informative priors centered at zero (no impact of the intervention):

$$\begin{aligned} \log RR &\sim \mathcal{N}(0, 10^2) \\ \sigma^2_{\log RR} &\sim U[0, 10]. \end{aligned} \tag{10}$$

3.3 Results

The results for Bayesian models with fixed and random effects are reported in rows 1 and 3 of Tables **??** (log scale) and **??** (raw scale) and in Figure **??**. For the BHM of all studies we find the mean RR is 0.75 (95% interval from 0.60 to 0.89), compared to 0.88 under FE model (95% interval from 0.83 to 0.93). Our FE result is identical to the value reported by Imdad et al (2022), whereas for RE we find a marginally higher benefits than the original meta-analysis.[10] The effect under RE model is more diffuse than under the

[10]The difference is likely due to using BHM which, as we discussed earlier, tends to estimate higher heterogeneity than typical frequentist implementations of RE models in statistical software.

FE model, but both models suggest that in the analysed sample of studies vitamin A supplements have on average considerably reduced child mortality. As we discussed in detail, the FE method gives a higher and more precise estimate of RR because it does not take into account the heterogeneity across sites.

Examining the posteriors for individual studies (Figure **??**) shows that the effects are heterogeneous. This can also be seen based on summary statistics for the model: the average pooling factor is 0.34, suggesting that 66% of the observed variation in estimated treatment effects is due to genuine differences between effects across studies (this value is typically referred to as I^2). However, heterogeneity is not precisely estimated, something that is typical to many meta-analyses of that size: Bayesian posterior 95% interval for I^2 is (26%, 91%). This strongly suggests presence of some heterogeneity, but we cannot learn its extent based on the sample of 18 studies.

This heterogeneity is also reflected in the posterior predictive distribution of RR for the RE model. It has a mean of 0.78, with 95% interval ranging from 0.4 to 1.3. (For the FE model it is by definition the same as the average effect.) This indicates that while we can be confident in the vitamin A effects within the analysed sample of studies, posterior predictive probability of RR being below 1 is 88%.[11]

Uncovering substantial heterogeneity prompts the question of why the impact is different across settings. In principle this question can be answered using contextual variables that describe differences in both the study designs and local contexts. In the vitamin A literature, there is presumably substantial variation in baseline vitamin A deficiency and the public health infrastructure that supports the intervention. The design of the trials also varies, particularly in the delivery mechanism which was direct in most cases but done via regional health centers in Awasthi et al (2013), leading to lower compliance with the supplement regime. Additionally, only some of the trials were double-blinded, and the intensity of the monitoring of outcomes per child differed dramatically across studies.

Unfortunately, the original papers do not report sufficient information to permit the construction of quantitative variables suitable for statistical analysis. For example, many studies did not collect data on the control group rates of vitamin A deficiency, since observing cases of xerophthalmia and nightblindness in the children of the communities being studied is indicative of a severe enough deficiency to plausibly justify intervention. This rationale makes sense within each individual study, but has lead to a situation in which it is now impossible to determine how much heterogeneity is due to characteristics of studied populations versus study design. This serves as a reminder that collecting data on contextual variables in field trials is important for the scientific process.

[11] We note that due to assumed normality of distribution of true effects the upper end of the distribution has a somewhat nonsensical interpretation: for example, the model implies a 2.5% change of 30% increase in the risk of mortality in a new study. Since we know that vitamin A supplementation is safe, a more appropriate model could consider a non-symmetrical distribution, with more probability mass concentrated around no effect rather than a harmful effect. Note, however, that this consideration is relevant for decision makers who are risk averse. In our case we focus on behaviour of the mean effect and do not consider the decision problem.

3.4 Publication bias

Similarly to previous authors, we find signs of publication bias and small0study effects in the vitamin A supplementation and child mortality literature. The funnel plot (which mirrors Figure 5 in Imdad et al 2022) is Figure **??**. Visual inspections suggests some asymmetry. This is further supported by the Egger regression, where intercept is significant under the random-effects specification (two-sided $p = 0.0067$, `metafor` in R), and also under a fixed-effect specification (two-sided $p < 0.0001$). By contrast, Begg–Mazumdar's rank correlation test is not statistically significant, and a simple regression of effects on (co)variance likewise does not reject symmetry. We interpret the overall pattern as indicative of small-study effects while noting that such diagnostics have limited power with $K \approx 18$ and can be confounded by genuine effect modification.

We then quantify the potential impact of selection using the Andrews–Kasy selection model with a symmetric significance threshold at $|z| = 1.96$. For the full set of trials (including DEVTA), the frequentist maximum-likelihood fit yields an estimated relative publication probability of $\hat{\omega} = 0.17$ (SE 0.12), together with bias-adjusted random-effects parameters $\hat{\tau}' = -0.13$ (SE 0.07) and $\widehat{\sigma'_\tau} = 0.09$ (SE 0.07). This corresponds to a pooled 12% reduction in mortality risk. Excluding DEVTA leads to similar results. These estimates are reported in Tables **??** (log scale) and **??** (risk-ratio scale; rows 5–6).

Relative to the unadjusted Bayesian random-effects fit ($\tau = -0.29$), the selection adjustment attenuates the pooled effect and substantially lowers the estimated between-study heterogeneity, consistent with the hypothesis of lower probability of reporting of non-significant results.

Given the modest number of studies and the identification limits of two-region selection models, we treat these adjusted estimates as likely evidence of need to shrink the RE estimates rather than dispositive corrections on the magnitude of effects. To further examine this, we ran an additional Monte Carlo simulations on simulated data. Our goal was to assess whether the Andrews–Kasy estimator can recover parameters in settings like ours. We chose the same $K = 18$ and same set of standard errors as in Imdad et al 2022, but with effect sizes drawn from the distribution estimated by BHM and with selection determined by reaching significance. In the presence of strong selection, $\omega = 0.17$, the procedure recovered both τ' and ω reasonably well on average (with $\hat{\omega}$ right-skewed but median close to the truth), and the induced attenuation in the pooled effect was similar to what we find empirically. However, when bias is absent ($\omega = 1$) the estimator was unstable with wide intervals and frequent near-boundary estimates.

3.5 Robustness to exclusion of DEVTA trial

The debate over the design of the DEVTA trial and the resulting difference in compliance rates relative to other studies has lead some researchers to suggest that it should not be included in meta-analyses. Sommer, West and Martorell (2013) dispute Awasthi et al.'s (2013) claims about the effectiveness of the DEVTA trial's supplement delivery mechanism and patient compliance, finally commenting "At best, DEVTA is but one unorthodox

study, done in one remote population of one country." Given the disagreement about whether DEVTA really belongs with the other studies, it is debatable as to whether it should be included in the meta-analysis.

We report on the results of meta-analysis without the DEVTA trial in Table **??** and Figure **??**. While the FE model is highly sensitive to exclusion of DEVTA, the RE model is affected much less. This difference between FE and RE models is illustrated in Figure 2, which shows strong overlap in p.p.d.'s for RE, but not FE, models.

It is important to also note that the heterogeneity estimate is not driven by exclusion of DEVTA. Whereas for 18 trials we found mean I^2 of 66%, for the model of 17 trials we have $I^2 = 50\%$ (95% interval from 1% to 86%). Thus the estimate of heterogeneity is even less precise, but on average the model still supports our concluson of heterogeneous effects.

Our result suggests that the DEVTA result does not substantially alter the evidence on typical effect of vitamin A supplements available from other studies and thus should not be taken as evidence that previous studies "overestimated" the effect in any major sense. Excluding the DEVTA trial also doesn't meaningfully change the estimated heterogeneity or pooling factors (Table **??**, columns 2 and 4).[12]

The DEVTA result is a precision outlier, a type of result which, as we have shown in our simulations, may have a detrimental effect on the performance of the FE estimator in presence of heterogeneity. Precision outliers do not heavily impede the performance of BHMs because they are designed to detect heterogeneity and reweight the evidence accordingly. It is not the case that highly precise studies are "underweighted" or "overweighted" in any given method. Rather, when true treatment effect are heterogeneous any single estimate should not heavily influence the general estimate no matter how precise it is. While high precision indicates high information about the site k in which the effect was estimated, this site is still just a singular data point in the set of K data points. High precision does not necessarily indicate highly generalizable information.

3.6 Formal model selection using cross validation

While our simulations and the high estimated heterogeneity are perhaps sufficient justification to choose an RE model over an FE alternative, we can also quantify the differences between the models by using a leave-one-out cross-validation approach (LOO CV).

For models without DEVTA trial, the performance of RE and FE models is very similar (RE elpd of -13.7, compared to FE elpd of -13.6).[13] However, when DEVTA trial is included, RE model still performs similarly (elpd of -14) while the FE model does not fit

[12]This makes sense in the context of our simulations, since DEVTA is not a "typical" location outlier. Four other studies which we include in our analysis have found higher point estimates, associated with smaller positive effects and even negative effects on child mortality (Table **??**, leftmost column).

[13]In this case, i.e. when DEVTA trial is excluded, the choice between FE and RE models in this purely data-driven way is difficult. While the elpd value is not directly interpretable and 18 studies is not enough to make a decisive comparison, both models are evenly matched in terms of number of studies for which they offer a better out-of-sample prediction (nine studies each).

data anymore (elpd of -32.6). This can be best illustrated in Figure **??**, which once again visualises p.p.d.'s, but this time does so for 18 different models, each time leaving out one study and fitting meta-analysis model to the remaining 17. The impact of DEVTA trial is clearly visible, although we can see that there are also other studies for which the FE model is inadequate.

In summary, the LOO CV procedure justifies choice of an RE model regardless of whether the DEVTA trial is included in the set of meta-analysed studies.

4 Conclusion

The controversy in the vitamin A literature stems in part from disagreement about how to aggregate a heterogeneous trial record. Fixed–effects (FE) meta-analysis targets the precision-weighted average effect in the included studies, whereas random–effects (RE) meta-analysis treats study effects as draws from a distribution and targets its mean across settings. In this application, there are clear *a priori* reasons to expect genuine between-study heterogeneity—differences in dosing schedules, baseline vitamin A deficiency and under-five mortality, delivery and compliance, making RE a more promising choice.

However, a recurrent objection is that RE "conceals the reliability" of very large trials by giving undue weight to small trials. While both FE and RE use inverse-variance weights, RE adds a between-study variance term, which flattens weights as heterogeneity increases. This makes small studies relatively more influential than under FE, a feature that is appropriate when the goal is to generalise beyond the exact study mix. However, that feature can also interact with selective reporting if present.

The methodological contribution of this paper to resolving that debate is two-fold. First, we review the FE and RE frameworks and show that in absence of publication bias there are strong theoretical and empirical reasons to prefer the RE approach in this instance. We do this by testing the models on simulated data, examining normal and heavy-tailed (Student-t) data-generating processes and mixture distributions with both location outliers and "precision outliers," i.e., single studies that are far more precise than the rest. Although the mean treatment effect results for frequentist and Bayesian inference coincide very closely for the RE model, we use BHM in order to report posterior predictive distributions (to derive prediction intervals for a new study), quantify pooling using the Gelman–Pardoe factor (which allows us to treat heterogeneity parameter as uncertain), and use LOO-CV, which clearly shows that out-of-sample performance drops when FE model is used with the DEVTA trial included.

Second, we aim to address selective-reporting concerns.We screen for small-study effects using funnel plots and regression tests, then fit a selection model that allows for lower publication probability of non-significant results. Specifically, we implement the Andrews–Kasy estimator with a symmetric two-sided threshold at $|z| = 1.96$, and jointly estimate the selection parameter and the RE hyperparameters to provide bias-adjusted pooled and heterogeneity estimate.

Using the Imdad et al. (2022) dataset, After adjusting for selective reporting with