The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1506–1518
https://doi.org/10.1214/23-AOAS1846

AS TREATED ANALYSES OF CLUSTER RANDOMIZED TRIALS

By Ari I. F. Fogelson[1,a], Kirsten E. Landsiedel[2,b], Suzanne M. Dufault[3,c] and Nicholas P. Jewell[4,d]

[1]*Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine,* [a]*Ari.Fogelson@lshtm.ac.uk*

[2]*Division of Biostatistics, School of Public Health, University of California, Berkeley,* [b]*kirsten_landsiedel@berkeley.edu*

[3]*Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco,* [c]*Suzanne.Dufault@ucsf.edu*

[4]*Department of Medical Statistics, London School of Hygiene and Tropical Medicine,* [d]*Nicholas.Jewell@lshtm.ac.uk*

Test-negative designs have rapidly become an appealing approach to assess disease interventions when randomization is not feasible and specifically used to measure the effectiveness of vaccines in the field (*Vaccine* **31** (2013) 2165–2168). An innovative extension of the test-negative design was recently used to assess the impact of a mosquito intervention where the intervention was applied at a cluster level with cluster assignment chosen at random, the AWED (applying *Wolbachia* to eliminate dengue) trial. The primary analysis reported was intention-to-treat (ITT) (*Trials* **19** (2018) 302; *N. Engl. J. Med.* **384** (2021) 2177–2186). However, the level of uptake of the intervention on mosquitoes was routinely captured in all clusters over time, and, furthermore, participants' mobility across clusters was measured in the time immediately preceding the onset of symptoms (whether test-positive or test-negative). Combinations of these measurements provide proxies for the true exposure to the intervention, thereby permitting an "as treated" assessment. We consider the use of marginal generalized estimating equations (GEE) and conditional generalized inear mixed models (GLMM) to estimate as treated efficacy, contrasting both with the ITT. We illustrate the strengths and challenges of these methods in the context of the AWED trial, highlighting several ways that common approaches to analysis of clustered data can yield incorrect results that can in turn be obscured and compounded by limitations in routine software. In addition, we estimate a greater level of intervention efficacy than shown in the ITT analysis.

1. Introduction.

Cluster-randomized trials employing test-negative sampling (CR-TND) aim to evaluate the efficacy of community-level interventions in an efficient and cost-effective manner (Anders et al. (2018a), Jewell et al. (2019)). Similar to traditional cluster-randomized trials (CRTs), clusters are randomized to either receive a community-level intervention or to act as control. The intervention status for any individual is determined by cluster membership such that intention-to-treat analyses classify individuals who live within treated clusters as treated and individuals within control clusters as not treated. Unlike traditional CRTs that require the enrollment and intensive longitudinal surveillance of cluster cohorts, test-negative sampling utilizes existing surveillance systems to identify and enroll symptomatic health-care seeking patients. Once enrolled, individuals are tested for the disease of interest, and cluster membership is recorded. Those who test positive are classified as cases, and those who test negative as controls. The process of sampling and enrolling individuals over the study period thus resembles a variant of a case-control study design (technically closest to a case-cohort design).

Received January 2023; revised September 2023.

Key words and phrases. As treated, cluster randomized trial, GEE, GLMM, per-protocol, test-negative design, dengue, Wolbachia.

Considerable work has been done to highlight the key assumptions necessary for unbiased estimation of intervention efficacy in the setting of a test-negative design (Jackson and Nelson (2013), Sullivan, Tchetgen and Cowling (2016)). These assumptions, and their criticisms, have been reexamined for use in a CR-TND by Anders et al. (2018a). For convenience the primary assumptions required are reiterated in the Supplementary Material (Fogelson et al. (2024)).

In the following we focus on estimation of the true comparative population incidence rate of testing positive across the two arms, the intervention relative risk. With no covariates, inference of this relative risk can be based on a variety of estimation techniques including the aggregate marginal odds ratio. When covariates are of interest and there are a reasonable number of clusters, the intervention relative risk can also be estimated through classical clustered modeling approaches such as marginal GEEs (Jewell et al. (2019)). Individual-level GLMMs, with random effects varying across clusters, target a related but distinct population parameter, a cluster-specific relative risk, as discussed in further detail in Section 3. A case-only approach is described in Dufault and Jewell (2020), but validity depends crucially on randomization if there is differential health-care seeking behavior across arms.

In some cluster-randomized designs, individual participant data may be available that capture adherence to the intervention. In such cases a secondary as treated analysis is often of interest. We note here the distinction between an as treated analysis (where measures of adherence are available at the individual level) and a per protocol analysis, which excludes individuals who do not adhere to the full intervention for a variety of reasons; see Smith, Coffman and Hudgens (2021) and Shrier et al. (2014). An as treated approach is possible for the AWED trial, where data on both individual mobility and spread of Wolbachia-infected mosquitoes allow consideration of contamination of control clusters and less than perfect coverage for individuals residing in intervention clusters. A primary purpose of this paper is to illustrate the necessity of going beyond typical marginal and mixed models approaches, as naive regression structures in these techniques may provide misleading results. In particular, we implement modelling approaches recommended in Neuhaus and Kalbfleisch (1998) and Begg and Parides (2003) for modelling clustered data with covariates that vary within clusters. Apart from the novel design of this trial, this allows broader consideration of potential sources of bias in handling clustered data, including from widely-used models that do not yield the estimands they intend.

These modeling issues are motivated by the World Mosquito Program's balanced parallel-arm CR-TND trial (AWED) that was designed to evaluate the efficacy of *Wolbachia*-infected mosquito deployment in reducing the burden of symptomatic dengue transmission in Yogyakarta City, Indonesia. With its population of approximately 400,000 persons, Yogyakarta was divided into 24 contiguous clusters each measuring approximately 1 km^2 in size. Twelve of the clusters were randomly assigned to the intervention arm, which received releases of *Wolbachia*-infected mosquitoes. *Wolbachia* successfully transinfected in nonnative hosts such as *Aedes aegypti* mosquitoes, the primary vectors of dengue, has been shown to disrupt the transmission of dengue and other flaviviruses by minimizing virus replication within the vector (Johnson (2015)). The remaining 12 clusters were assigned as control clusters (Figure 1).

Individuals who seek care at community health clinics who present with symptoms compatible with the clinical case definition of dengue and consent to enroll in the trial are subjected to laboratory testing for dengue, which determines their test-positive (case) or test-negative (control) status. Figure 1 shows the locations of the clinics. For further details of the trial, see Anders et al. (2018a) and Jewell et al. (2019), which provide discussion on the key assumptions.

Cavany et al. (2021) use the AWED trial as a case study and demonstrate how: (i) human mobility, (ii) intervention dilution, and (iii) cluster size may result in conservatively-biased

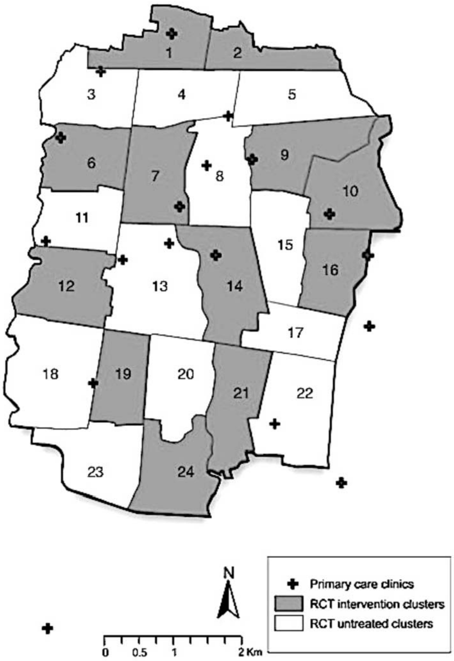


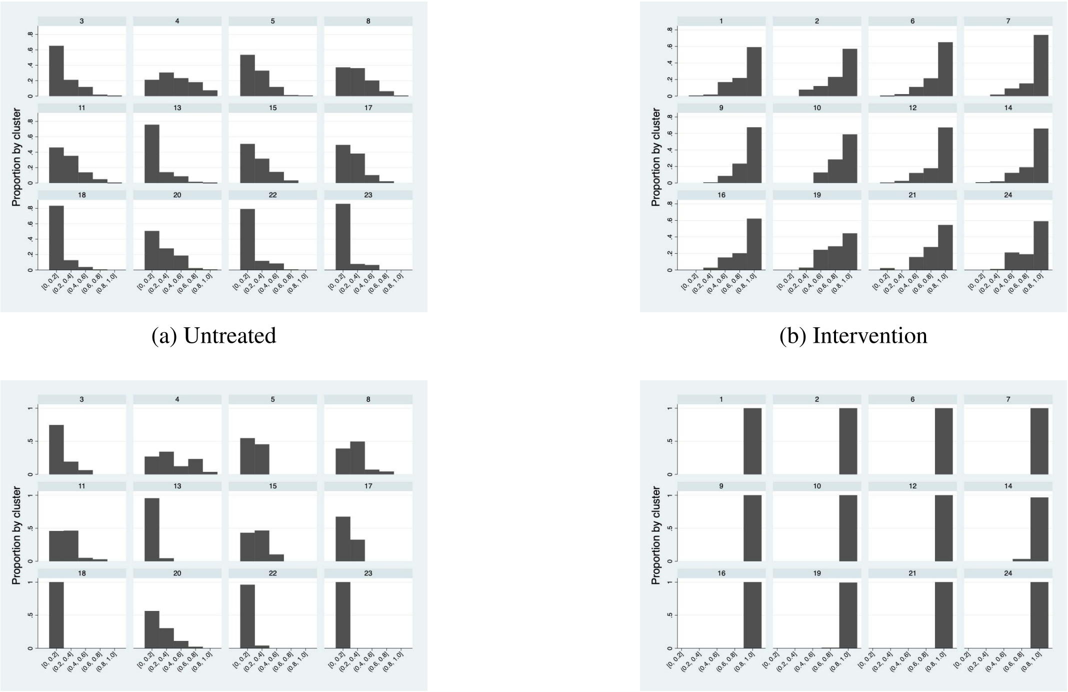
FIG. 1. *Map of Yogyakarta City trial area showing cluster assignments and locations of enrolling primary care clinics.*

efficacy estimates. Their discussion anticipates the broad thrust of our as treated assessment that examines the extent to which adjustment for human mobility and intervention dilution leads to greater efficacy estimates than initially reported.

2. Lack of perfect intervention in AWED trial. The AWED data allows for the calculation of two potential metrics for capturing true *Wolbachia* exposure at the individual-level (Anders et al. (2018b), Anders et al. (2020)). The first, "WEI Activity" (WEI-A) is based on: (i) human activity and (ii) *Wolbachia* prevalence in mosquitoes in all locations. The activity component takes into account self-reported locations where study participants spent time during daylight hours (5 a.m. to 9 p.m.) for the 10 days prior to reporting to a clinic with a febrile illness as well as the duration of the visits. In addition, for each cluster an aggregate *Wolbachia* prevalence was calculated monthly, based on mosquito trap surveillance data at several locations in each cluster. A continuous individual exposure index between 0 and 1 was then constructed by multiplying these cluster-level prevalence data during the month that the participant was recruited by their time spent at locations within any cluster. Calculation of individual exposure indices was carried out blinded to case/control status to remove observer bias. For further details, see Anders et al. (2018b).

An alternative exposure index was based solely on the individual's place of residence and measurements of the cluster-level *Wolbachia* prevalence in the participant's cluster of residence (for the calendar month of enrolment). We refer to this metric as "WEI Residence" (WEI-R). This metric ignores a participant's recent travel history, reflecting the possibility that dengue exposure risk is likely higher at home vs. other locations (Anders et al. (2018b)). Both exposure measures were predetermined in the study protocol (Anders et al. (2018b)).

Figure 2 shows crude histograms for observed WEI-A and WEI-R measurements for each participant (whether test-positive or test-negative) for each cluster, differentiated by intervention arm. WEI-A and WEI-R exposure indices for intervention and untreated clusters are clearly concentrated at higher and lower levels, respectively, more so for WEI-R in both arms.



(a) Untreated

(b) Intervention

(c) Untreated

(d) Intervention

FIG. 2. *WEI-A (top left and top right) and WEI-R (bottom left and bottom right) by cluster number, divided by intervention status.*

These phenomena are expected by design. The distribution of WEI-R in control clusters is influenced by some contamination by *Wolbachia*-infected mosquitoes over time, also reflected in WEI-A, the latter measure also capturing potentially impactful human mobility.

3. AWED as treated analyses using *Wolbachia* exposure indices. Initially, three simple models were fit to the data for both exposure indices that reflect common approaches to the modelling of clustered data: (i) a population-averaged logistic regression based on generalized estimating equations (GEE) methods, (ii) a logistic link generalized linear mixed model (GLMM) with random intercepts, and (iii) this GLMM supplemented by random slopes. For the GEE analysis we use both independent and exchangeable working correlation structures with robust standard errors. Under certain assumptions the choice of different working correlation structures may influence the efficiency of intervention effect estimates but not change the estimand of the estimates themselves. To illustrate methodological challenges, for simplicity we focus here on treating the exposure indices of the previous section as continuous explanatory variables with linearity assumptions on a logit scale for the infection outcome. Dufault et al. (2023) take a simpler approach using a categorical version of a closely related exposure index, thereby avoiding any linearity assumption. We return to these choices in the discussion.

Note that GLMM models yield cluster-specific effect estimates, while GEE methods provide marginal, or population-averaged, estimates that are directly comparable to previous ITT estimates. Due to the noncollapsibility of the odds ratio, these two methods target different population parameters. That is, the odds ratio conditioned on some factor (here cluster identity) differs from the marginal odds ratio obtained by collapsing over clusters (Daniel, Zhang and Farewell (2021)). In principal, we might anticipate the cluster-specific effect to be farther from the null than the marginal effect (Diggle et al. (2002)), assuming the former is based

TABLE 1
GEE results for WEI-A and WEI-R, odds scale

Marginal effects (GEE) for WEI-A and WEI-R		
	Working correlation	Odds ratio estimate (SE; p-value; 95% CI)
WEI-A	Independence	0.19 (0.06; <0.001; 0.11–0.34)
	Exchangeable	0.71 (0.20; 0.21; 0.41–1.22)
WEI-R	Independence	0.21 (0.05; <0.001; 0.13–0.35)
	Exchangeable	0.58 (0.23; 0.17; 0.26–1.26)

on a GLMM with random intercepts only. Regression estimates for both of these approaches are reported in Tables 1 and 2 for both WEI-A and WEI-R, where coefficients have been transformed to Odds Ratios for ease of interpretation.

3.1. *Naive model results.* The GEE estimate with an independent working correlation structure shows that exposure to *Wolbachia*, based on WEI-A, yields an estimated odds ratio of 0.19, reflecting an estimated reduction in the odds of contracting symptomatic dengue of 81%, when comparing no exposure (WEI-A = 0) to perfect exposure (WEI-A = 1), an effect that is highly statistically significant. This is very similar to the ITT estimated odds ratio of 0.23 (Utarini et al. (2021)) and to the as treated analysis provided in their Supplementary Figure S9A, which compared the highest WEI-A category from [0.8–1.0] to the lowest [0.0–0.2] yielding an estimated odds ratio of 0.75.

On the other hand, using an exchangeable working correlation structure, the estimated odds ratio is 0.71, or only a 29% reduction in risk associated with the intervention, without reaching statistical significance ($p = 0.21$). This is in marked contrast to both the ITT estimate and that achieved above with an independent working correlation. As anticipated, precision in estimation of the log-odds ratio (and thus the odds ratio) is somewhat greater when using exchangeable correlation. Similar results are found with the WEI-R index, although the difference between the point estimates using the two working correlation structures remains notable it is slightly more muted.

Of course, it is possible that the simple regression structure of Table 1 is misspecified, and we return to this below in Section 4. A more fundamental explanation arises from understanding the conditions required to obtain consistent GEE estimates that are robust to misspecification of the working correlation structure. Pepe and Anderson (1994) show that, in longitudinal settings, the selection of a working correlation structure may affect properties of estimates obtained by GEE methods. The issue occurs similarly with clustered data as here. Specifically, they note that a key condition for the consistency of GEE estimates is that a marginal expectation of the outcome equals a partly-conditional expectation,

$$\mathbb{E}(Y_{tj}|X_{tj}) = \mathbb{E}(Y_{tj}|X_{tj}, X_{ij}, i \neq t), \tag{1}$$

TABLE 2
Naïve conditional effects (GLMM) for WEI-A and WEI-R, odds scale

	Model	Odds ratio estimate (SE; p-value; 95% CI)	SD of slope (SE; 95% CI)
WEI-A	Random Int.	0.70 (0.24; 0.30; 0.35–1.38)	N/A
	Random Int. and Slope	0.63 (0.23; 0.21; 0.30–1.31)	0.68 (0.37; 0.23–1.99)
WEI-R	Random Int.	0.50 (0.23; 0.13; 0.20–1.22)	N/A
	Random Int. and Slope	0.98 (0.54; 0.97; 0.33–2.88)	1.77 (0.49; 1.03–3.06)

where Y and X are the outcome and covariate(s), respectively, and j indexes clusters and t and i individuals within a cluster.

For cluster-constant covariates, (1) is automatically satisfied. However, the modified score estimating equation used in GEE does not have mean 0, as required, if (1) fails to be true, except with a diagonal working correlation structure such as independence. In other words, when (1) fails, use of a nondiagonal structure, such as exchangeability, results in biased GEE regression estimates, even asymptotically. The literature often refers to this phenomenon as *GEE bias*.

The validity of (1) in the AWED trial is questionable since with an infectious agent such as the dengue virus, an individual's risk of infection may be influenced by other individuals' risk levels when those individuals are in close geographic proximity, such as living in the same cluster. This observation explains the similarity of the ITT estimate with the GEE effect based on an independence working correlation and, simultaneously, suggests that the exchangeable correlation regression results may be unreliable. We note that the rule of thumb of comparing naive and robust standard error (SE) estimates to assess adequacy of the working correlation structure does not identify the issue with exchangeability here. For example, for WEI-A, with an independent structure, the naive SE on the log odds scale is 0.17, compared to a robust SE of 0.30. With an exchangeable structure, the naive SE is 0.25, compared to a robust SE of 0.28, suggesting incorrectly that the exchangeable correlation working structure is reasonable.

We now turn to estimates based on a logistic GLMM which are, of course, not subject to GEE bias. We focus first on a simple random intercept model. The cluster-specific effect estimate for the same change in WEI-A (comparing WEI-A $= 0$ to WEI-A $= 1$) yields an estimated odds ratio of 0.70 (Table 2), or only a 30% reduction in the cluster-specific risk of infection, with a nonsignificant p-value of 0.3. Note that this cluster-specific estimate is naturally interpreted as the reduction in odds comparing no exposure (WEI-A $= 0$) to perfect exposure (WEI-A $= 1$) within any specific cluster, conditional on the random effect. However, this interpretation requires a leap of faith as most of the information for this estimate arises from between-cluster comparisons, an issue we explore in greater depth in Section 4. Similar results are found for WEI-R with an estimated cluster-specific odds ratio of 0.50, although again the difference between the marginal and GLMM estimated odds ratios are somewhat less pronounced. A $\bar{\chi}^2$ test for the variance of the random intercepts yields a p-value < 0.001 for both exposure indices. After accounting for the covariate, this test assesses whether there is evidence for statistical clustering of the outcome variable (here, dengue infection), thereby confirming the implicit suggestion of substantial within-cluster correlation from the GEE model (as shown by the need for a robust variance estimator). The $\bar{\chi}^2$ test is based on the likelihood ratio (LR) statistic whose asymptotic null distribution must be modified from standard LR methods since the null value (zero variance of the random intercepts) is on the boundary of the parameter space. The true asymptotic distribution involves a mixture of χ^2 distributions (Stram and Lee (1994)).

Based on an independent working structure, the estimated marginal effect of WEI-A is much farther from the null than the cluster-specific GLMM estimate, contrary to theory *if the random intercept model is correct*. This occurs also for WEI-R, although the difference is somewhat reduced. Some light is shed on this issue if one adds random slopes to the model. Neuhaus and Kalbfleisch (1998) showed that the property that cluster-specific effects are farther from the null than marginal effects does not hold, in general, in the presence of cluster variation in slopes. Here the estimated "average" odds ratio across clusters (obtained by exponentiating the average log-odds ratio) is 0.63 (Table 2), which is slightly closer to the estimated marginal effect although still considerably closer to the null. For WEI-R the "average" odds ratio across clusters is only 0.98, very close to the null but with a high estimated standard deviation for the slope distribution, presumably because there is so little variation in WEI-R in many of the clusters (particularly the intervention clusters). Use of either an Akaike or Bayesian Information Criterion indicates a slight preference towards the random intercepts only approach (we discuss more formal approaches to inference briefly below).

We make two observations: (i) there is little statistical evidence for the need for random slopes so that the latter only partially explain why the marginal effect is so much farther from the null, and (ii) the random slopes GLMM (for WEI-A) suggests that 25% of the cluster-specific slopes yield an odds ratio > 1, that is, a deleterious within-cluster effect of the intervention. The latter necessarily moves the average odds ratio closer to the null, thus empirically explaining part of the reason we see a cluster-specific effect that is less pronounced than the marginal effect. In this regard, what does one make of several clusters apparently reflecting a deleterious effect of exposure *within the cluster*? Individual logistic models, fit separately for each cluster, confirm this finding with more than half the clusters, 13 out of 24, suggesting a detrimental effect of the intervention based on the exposure index WEI-A. However, the individual cluster approach makes no use of information that compares infection patterns *across clusters*, the level at which the intervention is randomized.

In summary, irregularities in both the GEE and GLMM analyses suggest a more nuanced analysis of exposure through a decomposition of exposure effects within and between clusters, particularly in contexts like here where exposure varies far more between than within. We pursue this in the next section.

We end this section by noting some software challenges in conducting formal inference of the impact of adding random slopes to a random intercept logistic GLMM model. Stata 17 does not carry out $\bar{\chi}^2$ tests with the appropriate degrees of freedom to test additional random effects terms. A recently developed R package, vartestnlme, provides the capability to inferentially compare more complex models using the appropriate mixture distribution. However, common functions for fitting mixed models with binary outcomes in R, such as lmer and glmer, do not support adequate numerical integration estimates for models with random slopes—when adding random effect terms beyond random intercepts; these models are limited to a single adaptive Gaussian quadrature point, the Laplacian estimator, which is unsatisfactory.

4. Decomposing within- and between-cluster effects of *Wolbachia* exposure. To address concerns from the last section, we focus on two estimands simultaneously: (i) the effect of living in a cluster with a high level of Wolbachia as opposed to living in a cluster with a low level, the "between-cluster effect" (as captured by a covariate $\bar{X}_j$ that captures the mean exposure index for cluster j), and (ii) the effect of having greater or lesser Wolbachia exposure than the average for one's cluster of residence, the "within-cluster effect," as captured by the covariate X_{ij} for the ith subject in the jth cluster. These effects can either be estimated marginally (through GEE) or as cluster-specific effects based on a GLMM.

Neither of these cluster-specific estimands are the same as those obtained from the approaches of Section 3, which fail to differentiate these two effects. Neuhaus and Kalbfleisch (1998) call into question the conventional use of mixed models for clustered data based on solely modeling within-cluster-varying covariates. Coefficients from such models are typically treated as providing cluster-specific effect estimates, ignoring that between-cluster and within-cluster effects of a variable may be different, as is possible for the AWED intervention, where the protective effect of living in a high-Wolbachia cluster may be far greater than the impact of changes in exposure for an individual who lives in a protected area (or a control area for that matter).

This distinction is discussed further by Begg and Parides (2003) who argue that a key advantage of clustered data is the ability to assess these distinct estimands. In addition, for observational longitudinal data, the effect of an individual-constant covariate is likely subject to confounding by other individual characteristics (known or unknown), whereas the within-person effect is less likely to suffer such bias. We shall argue that the reverse is true here with clustered data, in part, because of the randomisation of the clusters to intervention.

Begg and Parides highlight several possible models that differentiate within- and between-cluster covariate effects, using what they refer to as cluster-level centering, and it is their judgment that the model proposed by Neuhaus and Kalbfleisch, using $\bar{X}_j$ for the between-cluster effect and $X_{ij} - \bar{X}_j$ for the within-cluster effect, should be avoided because of the likelihood of misinterpreting these parameters; they favor a model that simply uses the cluster-level mean exposure, $\bar{X}_j$, and the exposure of a single unit within the cluster, X_{ij}. Using the constructed variable $X_{ij} - \bar{X}_j$ leaves one open to misinterpreting the within-cluster effect.

We thus focus on the following model:

$$h(E[Y_{ij}|X_{ij}, \bar{X}_j]) = \beta_0 + \beta_W X_{ij} + \beta_B \bar{X}_j, \tag{2}$$

where h is the usual logistic link function. This represents a marginal model, but the same approach is immediately adaptable to the analogous GLMM. For example, a random-intercept GLMM may be specified as follows:

$$\begin{aligned} Y_{ij}|b_{0j} &\sim \text{Bernoulli}(\pi_{ij}), \\ \text{logit}(\pi_{ij}) &= \beta_0 + \beta_W X_{ij} + \beta_B \bar{X}_j + b_{0j}, \\ b_{0j} &\sim N(0, \sigma^2_{b_0}). \end{aligned}$$

Clusters are indexed by $j = 1, \ldots, n$, and individuals within clusters are indexed by $i = 1, \ldots, m_j$, where m_j is the number of individuals in cluster j. Y_{ij} is the dengue status of person i in cluster j, with their probability of dengue infection modeled as π_{ij}.

We assume here an understanding of the underlying assumptions of both the marginal and GLMM logistic regression models noted above. While the GEE marginal approach does not depend on any assumption about within-cluster correlation (after adjustment for covariates), regression model specification remains critical, as illustrated by our discussion of GEE bias above, as does the assumption of cluster independence. For GLMM models, model specification and cluster independence are also key assumptions while this approach further assumes conditional independence after adjustment for the random effects and included covariates; that is, the GLMM model assumes that within-cluster correlation can be entirely accounted for by including the appropriate random effects and covariates. For further discussion and examples of these two modeling approaches, see Begg and Parides (2003).

After decomposing the exposure indices in this fashion, new GEE and GLMM models were fit to assess and differentiate cluster-constant and within-cluster effects of *Wolbachia* exposure.

4.1. *Results.* From Table 3, using an independent working correlation, the decomposed GEE estimates of the marginal model (2) for WEI-A estimates the population averaged between-cluster effect of a one-unit increase in cluster-averaged *Wolbachia* exposure by an odds ratio of 0.06 (95% CI 0.02–0.14), holding an individual's exposure fixed. That is, this odds ratio reflects a comparison between two individuals who share identical WEI-A values but reside in either a cluster with all individuals having WEI-A = 1 vs. all having WEI-A = 0. Of course, in a finite sample it is not possible for such a comparison to exist in the observed data, so it is likely more interpretable to consider a smaller increase in the cluster average WEI-A (say, 0.5, e.g.) for the cluster-constant average exposure variable (with a 0.5 increase in $\bar{X}_i$, the estimated odds ratio is 0.24). This level of increase would marginally compare two

TABLE 3
Partitioned marginal effects (GEE) for WEI-A and WEI-R, odds scale

	Working correlation	Between-cluster coefficient (SE; p-value; 95% CI)	Within-cluster coefficient (SE; p-value; 95% CI)
WEI-A	Independence	0.06 (0.03; <0.001; 0.02–0.14)	1.56 (0.60; 0.25; 0.73–3.33)
	Exchangeable	0.07 (0.03; <0.001; 0.03–0.17)	1.54 (0.54; 0.23; 0.77–3.08)
WEI-R	Independence	0.02 (0.02; <0.001; 0.00–0.15)	7.17 (6.67; 0.03; 1.16–44.35)
	Exchangeable	0.03 (0.02; <0.001; 0.00–0.16)	6.88 (6.09; 0.03; 1.21–39.00)

individuals with the same individual level of WEI-A but who reside in two clusters whose average WEI-A differs by 0.5 units. The evidence for a positive effect of increased cluster-level exposure to the intervention is seen to be very strongly statistically significant ($p < 0.001$). This effect is estimated to be somewhat stronger for WEI-R.

From the same GEE model, the marginal within-cluster effect of a unit increase in WEI-A is captured by an estimated odds ratio of 1.56 (95% CI 0.73–3.33), holding the cluster-averaged exposure fixed. This marginal effect compares two individuals who reside in clusters with the same cluster average WEI-A (e.g., the same cluster) but whose individual WEI-A exposure differs by one unit. Again, it might be advisable to report the odds ratio associated with a smaller more realistic increase, smaller than one unit. This effect is seen to not be statistically significant, although it qualitatively suggests a surprising negative effect of exposure within clusters. We again see a similar effect for WEI-R with a very high odds ratio, albeit with also a very high standard error. The very substantial uncertainty in estimating this effect reflects far smaller within-cluster variation in WEI-R than WEI-A.

Using an exchangeable working correlation structure within GEE for WEI-A yields very similar findings with the marginal between-cluster odds ratio of 0.07 (95% CI 0.03–0.17) and within-cluster odds ratio of 1.54 (95% CI 0.77–3.08). It appears that condition (1) for the absence of GEE bias may likely be satisfied when one includes both individual and cluster average exposure measurements as covariates as shown in (2). The interpretation is then that an individual's risk of infection may be impacted by the average level of exposure in fellow cluster members but not by any additional pattern of those exposures so that, in this case, there is no longer a concern about GEE bias. Thus, this example illustrates that GEE bias can be reduced or eliminated by careful choice of covariates. This phenomenon is also noted when WEI-R is used.

Turning to the random intercept GLMM model, the conditional between-cluster effect of a one-unit increase in WEI-A is associated with an estimated odds ratio of 0.06 (95% CI 0.03–0.14) or an efficacy of 94% (95% CI 86%–97%), holding individual exposure fixed; see Table 4. This result is again highly statistically significant. The estimate for the within-cluster effect of a one-unit increase in WEI-A, holding the cluster average exposure fixed, yields an odds ratio of 1.58 (95% CI 0.74–3.34). While this result is not statistically significant ($p = 0.24$), it again suggests a detrimental within-cluster effect of additional *Wolbachia* exposure. For both estimates the confidence intervals are Normal-based but reflect a bootstrap estimate of the standard errors. Overall, a likelihood ratio test with a null hypothesis of equal between- and within-cluster exposure effects shows strong evidence in favor of the modeling approach with decomposed effects ($p < 0.001$). An additional likelihood ratio test, comparing the random intercept model with a model for WEI-A with a (i) a random slope for the within-cluster exposure variable, shows no evidence of heterogeneous slopes across clusters ($p = 0.91$). For WEI-R we also see similar results from the GLMM analysis as those produced by the marginal approach, as shown in Table 4.

TABLE 4

Partitioned Conditional Effects (*GLMM*) *for WEI-A and WEI-R*, *odds scale* (*the degrees of freedom for both models is four*)

	Model	Between-cluster coefficient (SE; p-value; 95% CI)	Within-cluster coefficient (SE; p-value; 95% CI)	AIC	BIC
WEI-A	Random Int.	0.06 (0.03; <0.001; 0.03–0.14)	1.58 (0.60; 0.24; 0.74–3.34)	2721.5	2748
WEI-R	Random Int.	0.02 (0.02; <0.001; 0.00–0.16)	8.15 (8.56; 0.05; 1.04–63.81)	2708.6	2735

With this choice of covariates, the conundrum about the sizes of cluster-specific vs. marginal effects has essentially disappeared. The cluster-specific and marginal effects are almost the same, with the former being very slightly farther from the null than the latter for both exposure indices. This suggests the the clusters do not vary substantially in their overall levels of dengue incidence (after adjusting for covariates), although random effects are necessary to capture within-cluster dependence.

For interpretation we note that, in either the marginal or GLMM model, the striking between-cluster effect is not likely to suffer from confounding by other cluster-constant covariates since average exposure is largely determined by the intervention status of cluster of residence, which was randomly allocated (and moreover, used constrained randomization due to the relatively small number of clusters). This view is supported by a comparison of baseline cluster and individual characteristics across the intervention arms, as displayed in Table S.1 in Utarini et al. (2021). However, within-cluster variation in individual exposure is observational and not randomly determined and thus may suffer from factors that determine why some individuals in intervention clusters have lower exposure indices with a similar comment for control clusters. For example, more affluent individuals (and their children) may have greater mobility (and thus lower exposures in intervention clusters) but be able to have higher protection against mosquitoes in their place of residence. In this regard, we note, however, that the unexpected positive association of within-cluster exposure and increased dengue risk does not appear to differ between intervention and control clusters (as measured by using an interaction term between exposure and within-cluster exposure in marginal or GLMM models).

An additional source of temporal confounding may arise because of the way individual deviations in exposure are calculated within a cluster. If we consider that *Wolbachia* prevalence is likely to only move in the direction of increasing prevalence, a month-to-month change in cluster-level *Wolbachia* necessitates that earlier recruits will tend to have lower exposure indices. It is also likely that variation will differ in the untreated and intervention clusters, with untreated clusters having larger swings in aggregate *Wolbachia* prevalence (as they begin at zero and slowly increase throughout the study period). There is considerable evidence from the study data that supports this speculation, as illustrated by Figure 2 of Utarini et al. (2021), in addition to increased dengue cases over the course of the study. To address the effect of calendar time, all observations were assigned to four calendar time periods that cover the 26 months of the trial (the first three of six months length, and the final one of eight months). Using indicator variables to include these time periods into the regression models produces a substantial change in the within-cluster effect for both WEI-A and WEI-R. Specifically, the GEE estimate of the marginal within-cluster odds ratio of a unit increase in WEI-A changes from 1.54 to 1.02, whether independent or exchangeable correlation is used, with the former having a 95% confidence interval of (0.55, 1.89), Similarly, the GLMM odds ratio estimate changes from 1.58 to 1.00 (95% CI: 0.49, 2.04). The estimated between-cluster odds ratio remains very low, at 0.09 (95% CI: 0.04, 1.18) under GEE with an independent working correlation (with very similar results under exchangeable correlation). The between-cluster GLMM estimated odds ratio is 0.10 (95% CI: 0.04, 0.26).

Thus, the adjustment for calendar time has completely eradicated any suggestion of a within-cluster effect of WEI-A. The effect of this time confounding is similar for WEI-R where the GLMM odds ratio is now estimated at the lower value of 3.82 (95% CI: 0.86, 16.87) with a p-value of 0.08. The marginal odds ratio estimates are similarly moved considerably closer to the null than before, and slightly smaller than the GLMM estimate, but with larger p-values that are greater than 0.3. In summary, it appears that any apparent impact of within-cluster variation of exposure is no longer evident when the confounding effects of time are removed.

We note that the exposure index, WEI-R, has no variation within a cluster over short time windows by definition, contributing to the lack of precision of estimated effects using this exposure measure. A cluster-averaged WEI-R also averages out variation over the course of the study, for example, due to contamination, introducing information from other time points with little relevance to risk of infection at a given time.

Finally, Neuhaus and Kalbfleisch (1998) provide further discussion of how regression coefficients from the naive models of Section 3 relate to the decomposed coefficients β_B and β_W of Section 4, as determined by characteristics of the covariate distribution within cluster and how such vary from cluster to cluster.

5. Discussion.

The models show very high conditional and marginal between-cluster efficacy when considering increases in either WEI-A or WEI-R. For WEI-A, at a cluster level, the intervention has very high efficacy. After addressing calendar time effects, there is little to no evidence of a within-cluster effect for individual exposure changes. The results demonstrate that the intervention is especially efficacious in cases where human mobility does not result in significantly reduced exposure to *Wolbachia*-infected *Aedes Aegypti*, yielding notably higher efficacy estimates than the ITT analysis. Specifically, efficacy is now always estimated to be greater than 90%, as compared to the reported ITT estimate of 77%.

Further consideration of the assumption of linear effects on the log-odds scale for exposure variables may be warranted, as the dose response may well not be linear across its full range. Given that many of the exposure levels are concentrated where we might anticipate a reduced dose response and that estimates of the between-cluster effects from the decomposed model are so close to 100% efficacy, we might expect to see both a diminished dose response and suspect extrapolation to larger changes of individual exposure. The latter reflects that we do not observe large within-cluster changes in exposure in either intervention or control clusters. This extrapolation also casts further doubt on the within-cluster exposure effects and explains the large uncertainty around such estimates. Dufault et al. (2023) avoid a linear assumption by using exposure categories, obtaining consistent results as ours for the between cluster efficacy (although they do not discuss the decomposition described in Section 4).

The value of the more nuanced measures of exposure have served a purpose in showing that the ITT estimate of the intervention efficacy is somewhat diluted, due to small deviations from perfect adherence to the intervention caused by mobility of both mosquitoes and participants. On the other hand, small deviations in individual estimated exposure for those living within the same cluster appear to make little difference to the risk of clinical dengue infection.

The exposure measures used here doubtless suffer from measurement error, which will only dilute intervention effects further. This adds to the attracation of exposure category comparisons, rather than relying on a continuous measure, as called for in the study protocol. We focused here on continuous measures, as this approach more simply illustrates the highlighted issues surrounding modelling; categorical analyses also yield similar findings, as noted.

Figure 2 illustrates extreme skewness of the exposure indices in either intervention or untreated clusters. This suggests that using mean cluster exposure to contrast with individual exposure in the decomposition of Section 4 may not be the optimal choice. For example, one could use the median cluster exposure as an alternative. Taking this issue further, other cluster-constant proxies of "average" cluster exposure might be suitable choice to "center" individual exposure, including cluster status (with regard to the intervention) itself. In the example both exposure indices are highly correlated with intervention status (see Figure 2), so this choice may be moot. But the general question deserves further research and illustration in other examples.

Statistically, these analyses illustrate that assessment of as treated effects in cluster studies is challenging and requires nuanced approaches. Naive applications of GEE of GLMM methods can lead to very misleading findings. The potential for GEE bias is substantial but can be reduced, or eliminated, by a careful choice of suitable covariates when available. This approach may allow an investigator to exploit potential precision gains by using nondiagonal working correlation structures without paying a price in bias. Further, decomposition of within- and between-cluster effects will generally be essential in understanding the true impacts of an intervention or exposure. Finally, an alternative intriguing causal inference approach to imperfect intervention adherence in cluster-randomized trials involves the use of the randomization indicator as an instrumental variable; for an introduction to these ideas; see, for example, Agbla, De Stavola and DiazOrdaz (2020), Agbla and DiazOrdaz (2018), and Kang and Keele (2018).

Acknowledgments. We acknowledge with deep appreciation the contribution of all investigators in the AWED Study Group to the implementation of the trial and collection of trial data and to the Tahija Foundation as funders of the AWED trial.

Funding. The second, third, and final authors were supported by grant #R01-AI148127 from the National Institute for Allergy and Infectious Diseases.

SUPPLEMENTARY MATERIAL

Supplement to "As treated analyses of cluster randomized trials" (DOI: 10.1214/ 23-AOAS1846SUPP; .pdf). Test-negative design assumptions; additional tables.

REFERENCES

AGBLA, S. C., DE STAVOLA, B. and DIAZORDAZ, K. (2020). Estimating cluster-level local average treatment effects in cluster randomised trials with non-adherence. *Stat. Methods Med. Res.* **29** 911–933. MR4078257 https://doi.org/10.1177/0962280219849613

AGBLA, S. C. and DIAZORDAZ, K. (2018). Reporting non-adherence in cluster randomised trials: A systematic review. *Clin. Trials* **15** 294–304. https://doi.org/10.1177/1740774518761666

ANDERS, K. L., CUTCHER, Z., KLEINSCHMIDT, I., DONNELLY, C. A., FERGUSON, N. M., INDRIANI, C., RYAN, P. A., O'NEILL, S. L., JEWELL, N. P. et al. (2018a). Cluster-randomized test-negative design trials: A novel and efficient method to assess the efficacy of community-level Dengue interventions. *Amer. J. Epidemiol.* **187** 2021–2028. https://doi.org/10.1093/AJE/KWY099

ANDERS, K. L., INDRIANI, C., AHMAD, R. A., TANTOWIJOYO, W., ARGUNI, E., ANDARI, B., JEWELL, N. P., DUFAULT, S. M., RYAN, P. A. et al. (2020). Update to the AWED (Applying Wolbachia to Eliminate Dengue) trial study protocol: A cluster randomised controlled trial in Yogyakarta, Indonesia. *Trials* **21** 429. https://doi.org/10.1186/s13063-020-04367-2

ANDERS, K. L., INDRIANI, C., AHMAD, R. A., TANTOWIJOYO, W., ARGUNI, E., ANDARI, B., JEWELL, N. P., RANCES, E., O'NEILL, S. L. et al. (2018b). The AWED trial (Applying Wolbachia to Eliminate Dengue) to assess the efficacy of Wolbachia-infected mosquito deployments to reduce dengue incidence in Yogyakarta, Indonesia: Study protocol for a cluster randomised controlled trial. *Trials* **19** 302. https://doi.org/10.1186/S13063-018-2670-Z

BEGG, M. D. and PARIDES, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat. Med.* **22** 2591–2602. https://doi.org/10.1002/SIM.1524

CAVANY, S. M., HUBER, J. H., WIELER, A., ELLIOTT, M., TRAN, Q. M., ESPAÑA, G., MOORE, S. M. and PERKINS, T. A. (2021). Ignoring transmission dynamics leads to underestimation of the impact of a novel intervention against mosquito-borne disease. MedRxiv 2021.11.19.21266602. https://doi.org/10.1101/2021.11.19.21266602

DANIEL, R., ZHANG, J. and FAREWELL, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom. J.* **63** 528–557. MR4226593 https://doi.org/10.1002/bimj.201900297

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

DUFAULT, S. M. and JEWELL, N. P. (2020). Analysis of counts for cluster randomized trials: Negative controls and test-negative designs. *Stat. Med.* **39** 1429–1439. MR4098500 https://doi.org/10.1002/sim.8488

DUFAULT, S. M., TANAMAS, S. K., INDRIANI, C., AHMAD, C. R. A., UTARINI, A., JEWELL, N. P., SIMMONS, C. P. and ANDERS, K. L. (2023). Reanalysis of cluster randomized trial data to account for exposure misclassification using a per-protocol and as-treated approach. Submitted for Publication.

FOGELSON, A. I, LANDSIEDEL, K. E, DUFAULT, S. M and JEWELL, N. P (2024). Supplement to "As treated analyses of cluster randomized trials." https://doi.org/10.1214/23-AOAS1846SUPP

JACKSON, M. L. and NELSON, J. C. (2013). The test-negative design for estimating influenza vaccine effectiveness. *Vaccine* **31** 2165–2168. https://doi.org/10.1016/J.VACCINE.2013.02.053

JEWELL, N. P., DUFAULT, S., CUTCHER, Z., SIMMONS, C. P. and ANDERS, K. L. (2019). Analysis of cluster-randomized test-negative designs: Cluster-level methods. *Biostatistics* **20** 332–346. MR3922137 https://doi.org/10.1093/biostatistics/kxy005

JOHNSON, K. N. (2015). The impact of Wolbachia on virus infection in mosquitoes. *Viruses* **7** 5705–5717. https://doi.org/10.3390/V7112903

KANG, H. and KEELE, L. (2018). Estimation methods for cluster randomized trials with noncompliance: A study of a biometric smartcard payment system in India. https://doi.org/10.48550/arXiv.1805.03744

NEUHAUS, J. M. and KALBFLEISCH, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54** 638. https://doi.org/10.2307/3109770

PEPE, M. S. and ANDERSON, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simulation Comput.* **23** 939–951. https://doi.org/10.1080/03610919408813210

SHRIER, I., STEELE, R. J., VERHAGEN, E., HERBERT, R., RIDDELL, C. A. and KAUFMAN, J. S. (2014). Beyond intention to treat: What is the right question? *Clin. Trials* **11** 28–37. https://doi.org/10.1177/1740774513504151

SMITH, V. A., COFFMAN, C. J. and HUDGENS, M. G. (2021). Interpreting the results of intention-to-treat, per-protocol, and as-treated analyses of clinical trials. *JAMA* **326** 433–434. https://doi.org/10.1001/JAMA.2021.2825

STRAM, D. and LEE, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50** 1171–1177.

SULLIVAN, S. G., TCHETGEN, E. J. T. and COWLING, B. J. (2016). Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *Amer. J. Epidemiol.* **184** 345–353. https://doi.org/10.1093/AJE/KWW064

UTARINI, A., INDRIANI, C., AHMAD, R. A., TANTOWIJOYO, W., ARGUNI, E., ANSARI, M. R., SUPRIYATI, E., WARDANA, D. S., MEITIKA, Y. et al. (2021). Efficacy of Wolbachia-infected mosquito deployments for the control of Dengue. *N. Engl. J. Med.* **384** 2177–2186. https://doi.org/10.1056/NEJMOA2030243