LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

MSc Project Report

2020-2021

Wolbachia, Dengue Fever and Randomised Test-Negative Studies

Candidate number: 2005647

Page count: 50

Project length: Standard

Submitted in part fulfilment of the requirements for the degree of MSc Medical Statistics

October 2021

Acknowledgements

I would like to thank my supervisor Professor Nicholas Jewell for his advice and guidance throughout the project, for introducing me to the field of spatial statistics which was the basis of the project and for obtaining the AWED trial data for me despite having some initial issues.

I would also like to thank Suzanne Dufault the University of California, Berkeley for her advice throughout the project, answering my many questions and for being extremely helpful in providing some vital R code necessary to carry out the analyses undertaken.

I also thank Dr. Chris Jarvis for introducing me to his method of cluster reallocation and providing guidance on how spatial statistics works in R and general advice in how to be organised when undertaking research work.

Finally, I would like to thank my personal tutor Professor James Carpenter for his continuous support and motivation throughout the past year and for giving me the opportunity to do this MSc, funded through the NIHR fellowship.

Contents

List of Figures

List of Tables

Abstract

Background: Dengue is a mosquito-borne viral infection that has rapidly spread in many parts of the world with 40% of the world's population at risk and approximately 400 million cases and 22,000 deaths occurring as a result of dengue annually. With no set treatment available for dengue and vaccines only being available in certain countries, dengue is a major public health burden. The Applying *Wolbachia* to Prevent Dengue (AWED) trial is a test-negative cluster randomised trial that took place in Yogyakarta, Indonesia which used the *wMel* strain of the *Wolbachia* bacteria to infect *Aedes Aegypti* mosquito to disrupt the transmission of dengue with this intervention showing an overall efficacy of 77.1% (95% CI: 65.3%- 84.9%). The aim of this project is to use the AWED trial data to assess the spatio-temporal relationship of dengue transmission for the untreated and intervention groups using various methods.

Methods: An individual's proximity to a virologically confirmed dengue (VCD) case in terms of distance of residence and dates of illness onset were used to define different levels of binary exposure which were implemented in contingency tables, generalised linear mixed models and generalised estimating equations to assess the effect exposure and intervention have on the odds of being a VCD case. The tau statistic was used for different levels of proximity between pairs of individuals to identify and quantify the existence of spatial dependence in dengue transmission in both the intervention and untreated clusters. Finally, the method of cluster reallocation was used to identify the existence of spatial spillover effects in the intervention and untreated clusters.

Results: There was strong evidence of an association between exposure and risk of being a VCD case, adjusting for intervention status and interaction. This decreased as distance between an individual and VCD case increased and also as difference in time of illness onset increased but only for shorter distances. There was evidence of spatial dependence in dengue transmission amongst those in the untreated clusters, but no evidence of spatial dependence for those in the intervention clusters which held true across differences of illness onset of 7, 14, and 30 days. There was evidence of spatial spillover in the untreated clusters and very little to no evidence of spatial spillover in the intervention clusters.

Conclusion: The introduction of *wMel*-infected *Aedes Aegypti* mosquitoes as the intervention in the AWED trial has shown to disrupt the spatio-temporal dependence of dengue transmission with there being little spillover effect in these clusters. There is a stronger relationship in dengue transmission spatially than temporally as much larger differences in estimates were observed for different distance proximities to a VCD case than the difference in estimates for different illness onset dates to a VCD case.

1. **Introduction**

1.1. Dengue

Dengue is a mosquito-borne viral infection that has rapidly spread in many parts of the world (1). Approximately 400 million cases and 22,000 deaths occur due to dengue worldwide annually (2). The virus responsible for causing dengue is the flavivirus, dengue virus (DENV). This flavivirus is part of the Flaviviridae family which is a positive stranded RNA containing virus; other viruses in this family include the yellow fever virus, West Nile virus, and Japanese encephalitis virus (2). There are four antigenically distinct, but closely related, serotypes of the dengue virus appropriately termed DENV-1, DENV-2, DENV-3 and DENV-4 (3). Being infected with one of these four dengue serotypes results in lifelong immunity against the specific serotype; however, cross-immunity against other serotypes after recovery is only temporary, meaning it is possible to be infected with dengue up to 4 times in one lifetime. Furthermore, subsequent infections of other serotypes increases the risk of developing severe dengue (3).

Dengue is typically found in urban and semi-urban areas of tropical and sub-tropical climates with the most significant dengue epidemics in recent years occurring in Southeast Asia, the Americas and the Western Pacific (4). The five countries reporting the most cases of dengue in 2020 were Brazil, Paraguay, Mexico, Vietnam and Malaysia (5). Approximately 141 countries are affected by dengue and around 40% of the world's population is at risk (4). Figure 1 highlights the severity of dengue cases worldwide for the year 2020.

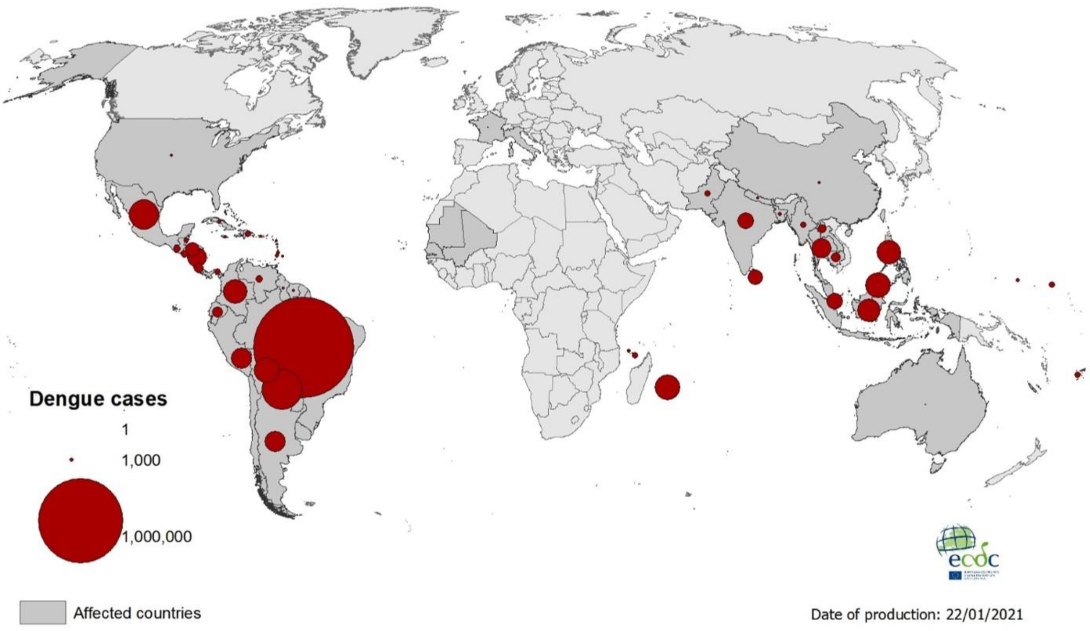

Figure 1 : Geographical Distribution of Reported Dengue Cases in 2020 (5)

1.2. Dengue Symptoms and Severe Dengue

Dengue virus infection introduces a wide variety of clinical symptoms ranging from mild fever to severe physiological conditions. Initial infection with a particular serotype of dengue typically results in mild disease manifestations known as dengue fever (2). These symptoms usually develop suddenly around 5 to 8 days after being infected with the virus (1). Dengue fever is characterised by symptoms such as high temperature, severe headache, muscle and joint pain, retro-orbital pain, and a widespread red rash (1). Dengue fever presents itself in three distinct phases: febrile, critical and convalescent (6). In the febrile phase, the fever typically lasts for 2-7 days and can be biphasic. The signs for this phase include the symptoms described above, as well as bleeding gums and epistaxis (6). Warning signs of progression to severe dengue occur in the late febrile phase at the time of defervescence and includes persistent vomiting, severe abdominal pain and difficulty breathing. The critical phase begins at defervescence and usually lasts for 1-2 days. Here, most patients improve clinically. However, some patients can experience severe haemorrhagic manifestations with the potential to lead to hepatitis or pancreatitis. Furthermore, patients experiencing substantial plasma leakage can develop severe dengue within a few hours as a result of increase in vascular permeability (6). Finally, the convalescent phase is deemed as the recovery phase in which the patient’s wellbeing improves and their white blood cell count usually starts to rise, along with a recovery of a platelet count. However, rashes and itching are also observed in this phase (2, 6).

Dengue hemorrhagic fever (DHF), also known as severe dengue, is a severe febrile disease characterised by haemostasis malfunction, increased vascular permeability, and severe increased vascular leakage that could lead to dengue shock syndrome. Dengue shock syndrome (DSS) is a form of hypovolemic shock that causes reduced peripheral perfusion and has the potential to lead to tissue injury and multi-organ failure (2). Approximately 1 in 20 patients with dengue virus disease progress to develop severe dengue (6).

1.3. Diagnosis of Dengue

There are several methods in which dengue can be appropriately diagnosed. Despite dengue fever and severe dengue possessing the numerous symptoms described above, approximately up to 40% - 80% of all dengue infections are asymptomatic (7). Methods for diagnosis for DENV infection include virological tests (that directly detects elements of the virus) and serological tests (which detect human-derived immune components that are produced in response to the virus) (3). Decisions regarding which diagnostic method should be used depends on the time of patient presentation to the laboratory or clinic.

Patient samples collected during the first week of illness should be tested by both serological and virological methods (3). DENV is found in serum, plasma or circulating blood cells or tissue in the first 1 to 7 days, usually during the period of fever (2). The virological method entails viral RNA being isolated for detection within this 1-to-7-day period by reverse transcriptase real-time polymerase-chain-reaction (RT-Q-PCR) amplification or by conventional PCR (2). Quantitative PCR (Q-PCR) can be used to quantify the viral load in bodily fluids (2). In general, virological methods are sensitive, but they require specialised laboratory equipment and technical training for staff implementing the test and are therefore not always available in all medical facilities (3). Serological detection depends on the demonstration of anti-dengue immunoglobulin M (IgM) antibodies or by non-structural protein 1 (NS-1) antigen in the serum or plasma of patients using either enzyme-linked immunosorbent assay (ELISA) or immune chromatographic-based rapid card tests (2). IgM antibodies are detectable approximately 1 week after infection and are highest at 2 to 4 weeks after the onset of illness; they remain detectable for around 3 months after illness onset (3). The NS-1 and IgM diagnostics are not fully reliable due to cross-reactivity with other flaviviruses such as the Zika virus. Some other serological tests that are more accurate in diagnosing dengue infection include a neutralisation test and hemagglutination-inhibition (2).

1.4. Treatments for Dengue

There are currently no universal treatments for patients that are infected with dengue. Taking general painkillers or fever reducers such as paracetamol can help to relieve muscle aches and pains (1). However, non-steroidal anti-inflammatory drugs such as aspirin and ibuprofen should be avoided as these drugs act by thinning the blood which may exacerbate the prognosis for diseases with risk of hemorrhage (3). Furthermore, medical care and maintenance of the patient’s body fluid volume is critical for treating severe dengue as it carries a much higher risk than dengue fever (4). However, there have been studies on several sulphated polysaccharides extracted from seaweeds in which high antiviral activity against dengue have been observed (2).

It is evident that dengue carries a large burden globally and the need for a vaccine against dengue is of paramount importance. The first dengue vaccine, Dengvaxia (CYD-TDV), developed by Sanofi Pasteur was licensed in December 2015 and has now been approved by regulatory authorities in around 20 countries (3). This vaccine was shown to be efficacious and safe for individuals who have had previous dengue infection (seropositive individuals). However, this vaccine also has an increased risk of severe dengue for individuals who experience their first natural dengue infection after being vaccinated (those who were seronegative at the time of vaccination) (3). It is therefore targeted for individuals living in endemic areas ranging from 9 – 45 years old, who have had at least 1 documented dengue virus infection previously (3). Hence, if other countries want to implement CYD-TDV as part of their dengue control programme, pre-vaccination screening would be a recommended strategy. This would mean that only those with evidence of previous infection, either through an antibody test or a documented laboratory confirmed dengue infection, would be vaccinated (3).

1.5. Vectors of Dengue

In epidemiology, a vector refers to a living organism that can transmit infectious pathogens between humans, or from animals to humans (8). Vectors are mostly comprised of bloodsucking insects that ingest disease-producing microorganisms during a blood meal from an infected host and thereby transmitting it to a new host after the pathogen has been replicated (8). Here, the primary and secondary vectors of the dengue virus are the female *Aedes Aegypti* mosquito and the lesser known *Aedes Albopictus* mosquito respectively (2). Although the *Aedes Aegypti* mosquito's official common name is the "yellow fever mosquito", it is predominantly a public health concern as the main vector of DENV due to there being an effective vaccine for yellow fever (9). It is posited with near certainty that the *Aedes Aegypti* originated from sub-Saharan Africa with the mosquito spreading across various continents via ships and this method of dispersal is thought to have the highest risk of introducing the mosquito to mainland Europe (9, 10). Currently, they are found throughout the tropics including Africa and numerous sub-tropical regions including the south-eastern United States, the Middle East, Southeast Asia, the Pacific and Indian islands, and Northern Australia (10). The *Aedes Aegypti* will preferentially feed on humans, even in the presence of other mammals (10). They are typically found within 100m of human habitations due to the host-seeking opportunities and have adapted to utilise both indoor and outdoor aquatic containers such as vases or water tanks in order to breed (2, 10). The *Aedes Albopictus*, commonly known as the "Asian tiger mosquito", is similar to the *Aedes Aegypti* both in appearance (both having a black and white pattern on their bodies) and feeding habits as they also preferentially feed on humans in the daytime (11). However, the *Aedes Albopictus* is a more aggressive biter, mostly living outdoors where they are known for significantly reducing the quality of life in places such as Italy, southern France, and Spain (11).

2. Background, Aims and Objectives

2.1. Dengue Control Strategies

Since dengue is now endemic in more than 100 countries in the world (3), the urgency to discover effective prevention and control strategies to eliminate the virus is ever-increasing. There have been a number of strategies previously used to prevent dengue either directly through the use of a vaccine, or indirectly by reducing the prevalence of *Aedes Aegypti* or *Aedes Albopictus* mosquitos. Prevention and control strategies for reducing DENV indirectly are divided into three categories: Physical control, biological control, and chemical control (12). Some physical control strategies used in the past involve utilising Geographic Information System (GIS) mapping techniques to locate dengue concentrations and treating seropositive cases accordingly, and effective surveillance that enables the understanding of the spatiotemporal distribution of dengue cases to provide better planning (12). However, it was shown that surveillance alone was not sufficient after an outbreak of dengue occurred after decades of surveillance in Singapore in 2005 due to unsustainable vector control measures (12). There has also been some success through the use of community-based strategies such as educating local people in areas where dengue is present through media outlets in order for them to be able to identify and deal with vector habitats and use preventative measures (12). Chemical control strategies used previously involve the use of insecticides on the mosquitos, which is the most commonly used integration strategy (12). Although this method has been used for decades, the continuous use of insecticides can have a negative impact on the environment through the contamination of water, soil and can be toxic to other organisms such as birds or fish (13). Furthermore, it was shown that insecticides are not effective in reducing the prevalence of *Aedes Albopictus* due to increased resistance to insecticides as opposed to *Aedes Aegypti* (11, 14). Another chemical control strategy involves using pheromones as an “attract-and-kill” approach (12). A study involved developing an uncomplicated “lethal-lure control” based on this approach and found that by using pheromones and insecticides together, it attracted mosquitos and also restricted the hatching of eggs and killed the larvae, highlighting its usefulness (12). One biological control utilised the fact that *Aedes Aegypti* larvae reside in bodies of water by introducing larvivorous guppy fish into water containers. It is viewed as a cost-effective and eco-friendly strategy in controlling the population of *Aedes Aegypti* and was shown to successfully reduce the larval population in a study in Cambodia (12). The most promising biological control strategy is the genetic control of *Aedes Aegypti* in which paratransgenesis is the popular method (12). Paratransgenesis in general uses genetically-modified symbiotic bacteria that are reintroduced in the vector to colonise the vector population, therefore limiting the transmission of the disease in question (12). In this case, the most effective bacterial agent used in *Aedes Aegypti* to eliminate dengue is *Wolbachia* (12).

2.2. Wolbachia

Wolbachia pipientis is a common, maternally inherited bacteria that occurs naturally in 60% of insect species including some mosquitos, fruit flies, moths and dragonflies, but not naturally occurring in *Aedes Aegypti* (15, 16). *Wolbachia* live inside insect cells and are passed from through generations via an insect's eggs (15). It has been shown that *Wolbachia* are safe for humans, animals and the environment through multiple independent risk assessments which showed negligible risk associated with *Wolbachia*-infected mosquitos (15). Transinfection of *Aedes Aegypti* with some strains of *Wolbachia* grants resistance to infection by DENV and other arboviruses (16). Hence, the introgression of "virus-blocking" strains of *Wolbachia* into populations of aedes aegypti is an innovative dengue control measure (16). This approach entails regular releases of *Wolbachia*-infected mosquitos into a wild-mosquito population over a period of several months in which population introgression is achieved through manipulation of reproductive outcomes; the only viable mating option being that where all progeny are infected with *Wolbachia* (16).

2.3. Dengue in Indonesia

Indonesia is one of the largest countries in the DENV endemic region with a population of 251 million (17). Dengue has been prevalent in Indonesia for over 50 years with the first 58 cases being reported in Jakarta and Surabaya in 1968. Since then, the rate of incidence of severe dengue cases has increased over 700 times from around 0.05 cases per 100,000 person-years in 1968 to between 35-40 per 100,000 years in 2013 as shown in Figure 2 below (17). Hence, a clear understanding of the current epidemiology of DENV in Indonesia is critical for design of appropriate public health measures (18).

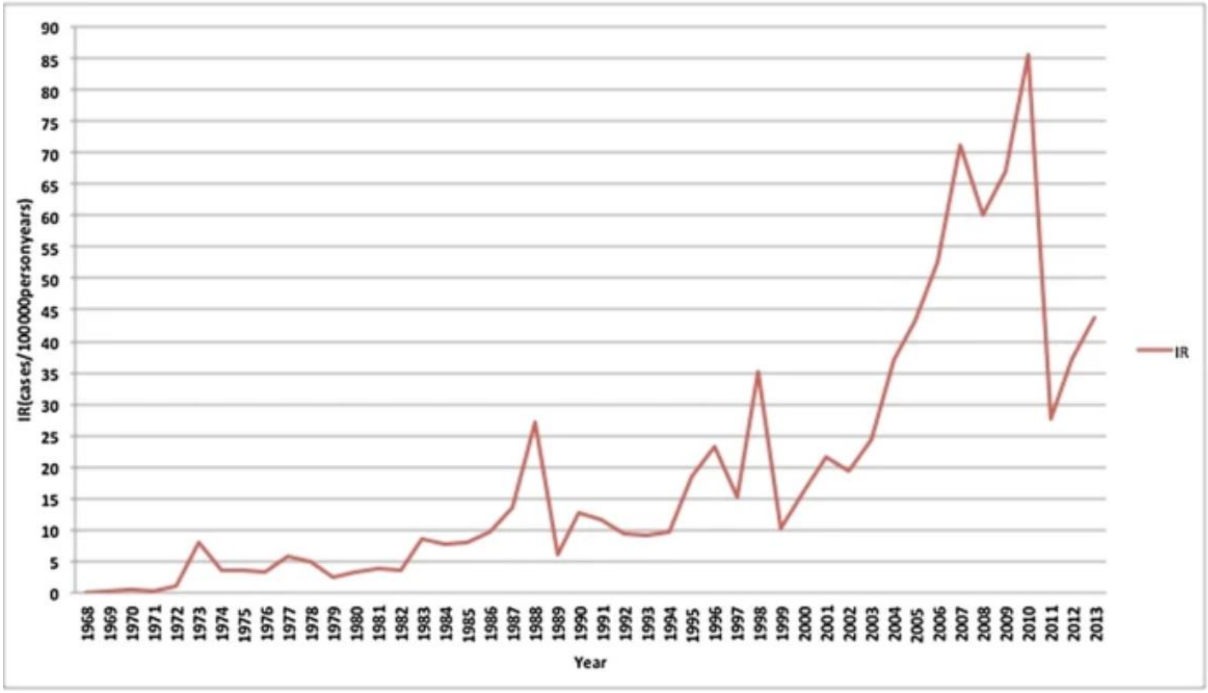

Figure 2: Trend in Incidence Rate (IR) of DHF cases in Indonesia from 1968-2013, measured in cases per 100,000 person-years (17)

2.4. AWED Trial

This *Wolbachia*-specific biological control strategy was successfully implemented in the Applying Wolbachia to Eliminate Dengue (AWED) trial; this trial is the basis for this project. The study was a cluster randomised controlled trial, designed to assess the efficacy of the deployments of *Aedes Aegypti* mosquitos infected with the *wMel* strain of *Wolbachia* in reducing the incidence of virologically confirmed dengue cases in Yogyakarta, Indonesia (16). The study area of Yogyakarta is a contiguous urban area spanning 26km^2, and has a population of approximately 311,700 (16). The trial design divided this study area into 24 clusters, each with approximate area of 1km^2 such that 12 clusters were selected at random to receive the intervention (the deployment of *wMel*-infected *Aedes Aegypti*) and the other 12 controls clusters receiving no deployment (using a form of constrained randomisation). The trial was loosely blinded in the sense that most community members in the intervention clusters were unaware of cluster assignment as the release containers were discretely placed in a minority of residential properties for a short period of time (16). Each intervention cluster received between 9 and 14 rounds of deployments of mosquitos eggs between March and December 2017 (16). Participants were subsequently recruited from a network of 18 government-run primary care clinics in Yogyakarta and the nearby Bantul district. Participants were eligible if they were 3 to 45 years of age, had a fever (either self-reported by participant or measured in the clinic defined by a forehead or axillary temperature of over 37.5°) with onset of 1 to 4 days beforehand, and had resided in the trial area every night for

10 days preceding illness onset (16). The exclusion criteria included participants having symptoms suggestive of a specific diagnosis other than an arboviral infection such as diarrhoea or pneumonia, or those that were enrolled in the trial within the 4 previous weeks (16). Trial participants were classified as being virologically confirmed dengue (VCD) cases using either virological or serological methods previously described. They were VCD if the plasma sample obtained at enrolment tested positive for DENV in a RT-Q-PCR assay or if the plasma sample tested positive for DENV NS-1 antigens in an ELISA (16). They were classified as test-negative controls if all diagnostic tests produced negative results (16). Furthermore, the specific serotype of DENV was determined through the use of a separate RT-PCR assay (16).

An interesting feature of this trial is that it utilises the test-negative study design. The test-negative design was developed as an efficient approach to assessing influenza vaccine effectiveness using available surveillance structures with the first publications using this design coming from Canada in 2005 (19). The test-negative design is a type of case-control design which, in the context of DENV, is one where intervention classification is compared between DENV test-positive cases versus DENV test-negative controls both of which present to a clinician with the pre-defined eligibility criteria of dengue-like illness (19). Classical case-control studies typically require intense efforts to recruit non-diseased controls, whereas the test-negative design (which also resembles the case-cohort design) recruits controls from the same source population as the cases being ill patients who are tested to identify a specific aetiology of interest for their illness (19). This highlights the relative cost-effectiveness and convenience of the test-negative design in comparison with the usual case-control design, but there are also concerns in the validity of this design due to the potential for misclassification of cases as controls due to imperfect test sensitivity (19).

The primary endpoint for the AWED trial was symptomatic VCD of any severity caused by any DENV serotype (16), at least severity sufficient to stimulate a clinic visit. The secondary endpoints were symptomatic VCD caused by each of the four serotypes, DENV-1, DENV-2, DENV-3, DENV-4, and symptomatic, virologically confirmed chikungunya and Zika virus infections (16). The minimum sample size required was 400 VCD cases and 1600 test-negative controls and this was calculated such that the trial had 80% power to detect a 50% lower incidence of VCD among participants in intervention clusters than among those in control clusters (16). Due to the emergence of SARS-CoV-2 in Yogyakarta in March 2020, enrolment for this trial was forced to end early meaning recruitment at the 18 primary care clinics lasted from January 8th, 2018, to March 18th, 2020 (16). As a result, only 385 VCD cases had been enrolled, although there were 5921 test-negative controls.

Wolbachia exposure was defined as a binary classification based on if the participant's primary place of residence was in an intervention or control cluster for 10 days prior to illness onset (16). The outcome was VCD with the intervention effect being estimated from an aggregate odds ratio comparing the exposure odds among participants with VCD with the exposure odds among the test-negative controls using the constrained permutation distribution as a basis for statistical inference (16). The null hypothesis was that the odds of residence in an intervention cluster would be the same among participants with VCD and those that are test-negative controls and efficacy was calculated as

$$100\ \times (1 - aggregate\ odds\ ratio)$$

The results from the AWED trial for the intention-to-treat analysis showed strong evidence (p=0.004) to reject the null hypothesis with an overall protective efficacy of 77.1% (95% CI: 65.3, 84.9) (16). This means that the odds of residing in an intervention cluster was 77.1% lower for participants with VCD than the odds of residing in an intervention cluster for test-negative control participants, highlighting the success of *wMel* deployment in the intervention clusters in reducing the incidence of DENV.

2.5. Aims and Objectives

The aim of this project is to conduct a secondary analysis of the AWED trial by:

- Assessing cluster-specific and population-averaged effects of exposure on dengue transmission, when exposure is defined in a spatio-temporal manner
- Implementing spatiotemporal methods to compare spatial clustering of dengue between intervention and untreated clusters and how it changes over distance and time
- Investigating the existence of a spillover effect in the intervention and untreated clusters

3. Generalised Linear Mixed Models and GEEs

3.1. Exploratory Analysis

The data set used for this project is a subset of the final closed data set from the AWED trial. It is a relatively simple dataset only containing 11 variables in total. These were participant ID, their age and sex, the cluster of primary residence for 10 days prior to illness onset, whether or not their cluster received deployment of *wMel*-infected *Aedes Aegypti* (i.e. intervention status), the latitude and longitude of their primary residence, dates of both illness onset and enrolment into the trial, binary dengue status, and finally the specific serotype. Participants that were VCD cases were classified as having an unknown serotype if they were diagnosed as negative through the RT-PCR test, but positive for DENV NS-1 antigens in an ELISA (16). A map of Yogyakarta highlighting the arrangement of clusters, their intervention allocations along with the locations of the primary care clinics is shown in Figure 3 below.

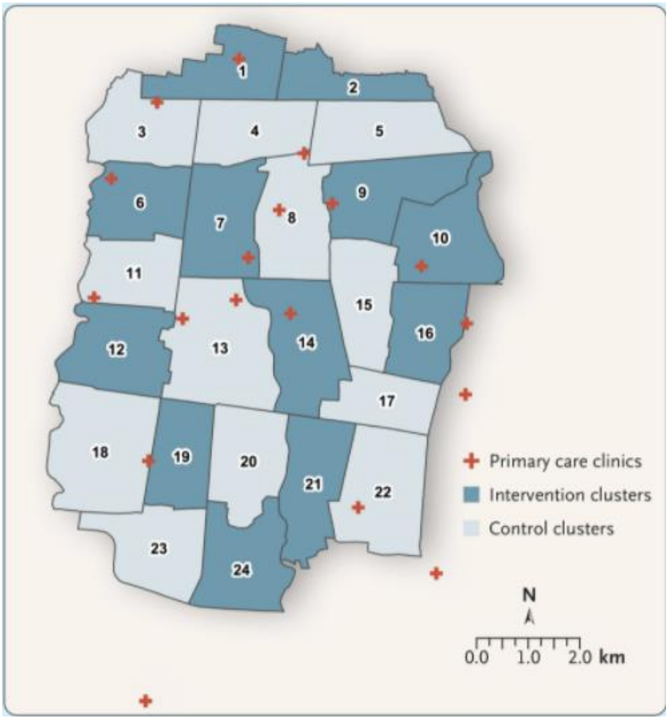

Figure 3: Map of Yogyakarta, Indonesia, showing clusters and intervention status with location of recruiting primary care clinics (16)

Note that all analyses for this project was done using R (version 4.1.0). As previously mentioned in Section 2.4, there were 6313 total participants in the AWED trial of which there were 392 (6.2%) recorded VCD cases and 5921 (93.8%) recorded test-negative controls. Out of these 392 VCD cases, the most commonly observed serotype was DENV2 with 154 (39.3%) cases and the least commonly observed serotype was DENV3 with only 27 (6.9%) cases. 7 of these recorded 392 VCD cases were the result of a second DENV infection during the study period, of which the serotype of the second infection was different to the serotype of the first due to the nature of the dengue virus. Hence, there are in fact 385 distinct participants that were VCD cases. 67 (17.4%) participants had a serotype that was unknown. Table 1 shows the distribution of participants and VCD cases across each cluster.

Cluster	N (%)	VCD Cases (%)	Mean Age (SD)	Sex (%)*	Cluster	N (%)	VCD Cases (%)	Mean Age (SD)	Sex (%)*
1	274 (4.4)	8 (2.9)	16.2 (10.5)	129 (47.1)	13	794 (12.6)	76 (9.6)	14.1 (9.4)	386 (48.6)
2	91 (1.4)	3 (3.3)	17.0 (10.0)	55 (60.4)	14	205 (3.2)	7 (3.4)	14.1 (9.0)	87 (42.4)
3	449 (7.1)	26 (5.8)	16.1 (11.6)	217 (48.3)	15	147 (2.3)	12 (8.2)	15.6 (9.1)	76 (51.7)
4	190 (3.0)	11 (5.8)	15.6 (10.7)	88 (46.3)	16	153 (2.4)	3 (2.0)	17.2 (10.0)	81 (52.9)
5	161 (2.6)	31 (19.3)	16.3 (8.9)	79 (49.1)	17	89 (1.4)	9 (10.1)	18.0 (10.1)	48 (53.9)
6	666 (10.5)	10 (1.5)	14.9 (10.5)	331 (49.7)	18	344 (5.4)	49 (14.2)	14.3 (10.0)	169 (49.1)
7	507 (8.0)	7 (1.4)	14.4 (9.6)	250 (49.3)	19	140 (2.2)	5 (3.6)	15.3 (9.0)	67 (47.9)
8	324 (5.1)	19 (5.9)	15.2 (10.4)	163 (50.3)	20	173 (2.7)	28 (16.2)	14.0 (9.9)	82 (47.4)
9	141 (2.2)	4 (2.8)	13.9 (8.8)	70 (49.6)	21	90 (1.4)	3 (3.3)	17.6 (11.8)	44 (48.9)
10	96 (1.5)	9 (9.4)	16.7 (10.0)	53 (55.2)	22	180 (2.8)	29 (16.1)	16.6 (10.7)	93 (51.7)
11	415 (6.6)	23 (5.5)	14.7 (10.1)	181 (43.6)	23	141 (2.2)	11 (7.8)	16.4 (10.9)	63 (44.7)
12	448 (7.1)	7 (1.6)	14.0 (9.6)	226 (50.4)	24	95 (1.5)	2 (2.1)	14.2 (10.0)	37 (38.9)

Table 1: Descriptive Statistics of Trial Participants in AWED Trial

Note: N = Number of Participants; SD = Standard Deviation; Age is measured in years; Sex column contains the number of females; shaded rows represent intervention clusters

The mean age of participants across each cluster appear to be quite similar with means varying from around 14 to 17 years of age. Cluster 17 had the highest mean participant age of 18.0 years old. The overall mean age of participants was 15.1 years old. However, there was quite a large spread of ages across all clusters with the smallest standard deviation being 8.8 years in cluster 9 and the largest being 11.8 years in cluster 21. There were also a lot of outliers observed for the upper bound of ages as 18 out of the 24 clusters contained at least 1 participant with an age that was an outlier for that cluster. This is highlighted in a boxplot that can be found in the appendix. The mean age for VCD cases is 13.1 years old whereas the mean age for test-negative controls is slightly older being 15.2 years old. The distribution of male and female participants was quite even across the majority of the clusters, though cluster 24 saw quite a low proportion of females with only 38.9% and cluster 2 saw a larger proportion of females with 60.4%. Similarly, the number of VCD cases were roughly equal amongst males and females with 201 (51.2%) cases being male and 191 (48.8%) cases being female. Cluster 13 had the most amount of VCD cases with 76 which is 19.4% of the total number of cases; this would be expected as this cluster also had a distinctly larger number of participants (794) compared to most other clusters. Cluster 5 had the highest proportion of cases with 19.3% of the participants in this cluster being a VCD case. Cluster 24 had the least amount of cases with only 2, and clusters 2, 16 and 21 each only had 3 cases. There was also no missing information from any participant.

3.2. Introduction to Model Building

3.2.1 Assigning Exposure Status

An important aspect about the analysis of the data for this project is that intervention status will not be used as the main exposure of interest, but instead we define a binary exposure variable in such a way that incorporates both spatial and temporal information received from the participants.

Let $d_1 = Lower\ bound\ for\ radius\ (metres), d_2 = Upper\ bound\ for\ radius\ (metres)$

Let $t = Time\ (days)$

Then we can define binary exposure, X_i, for participant i as

$$X_i = \begin{cases} 1\ if\ participant\ i\ is\ within\ [d_1, d_2)\ metres\ of\ a\ VCD\ case\ AND\ within\ t\ days \\ 0\ otherwise \end{cases}$$

This is achieved by first creating a $6313\ \times 6313$ indicator matrix for distance, D_{ij}, such that

$$D_{ij} = \begin{cases} 1\ if\ individual\ i\ is\ within\ [d_1, d_2)\ metres\ of\ individual\ j \\ 0\ otherwise \end{cases}$$

Note that since we are looking at individual i's exposure to a VCD case, this means individual i's illness onset must occur after that VCD case. Also, the upper bound is an open interval meaning individual i must be less than d_2 metres from individual j.This utilises the *geodist* package in R which uses the latitude and longitude coordinates to compute an interpretable distance. Secondly, another $6313\ \times 6313$ indicator matrix is created for time, T_{ij}, such that

$$T_{ij} = \begin{cases} 1\ if\ individual\ i\ has\ illness\ onset\ within\ t\ days\ of\ illness\ onset\ of\ individual\ j \\ 0\ otherwise \end{cases}$$

Multiplying matrices D_{ij} and T_{ij} produces another $6313\ \times 6313$ indicator matrix such that the cells are 1 if both conditions are true. Finally, exposure is assigned by identifying which individuals are VCD cases which is done by filtering the resulting matrix such that the columns only contain individuals who are VCD cases since j represents columns and we want to identify whether or not individual i has been in close proximity to a VCD case.

For example, suppose $d_1 = 0, d_2 = 200, t = 10$. Then we would say that a participant is exposed if they have been less than 200m from a VCD case within 10 days of their illness onset. However, they would not be exposed if they had been within 200m from a VCD case 14 days before their illness onset or if they had been 250m from a VCD case 5 days before their illness onset.

3.2.2 Contingency Tables

Recall from Section 2.4 that our outcome is Y_i which is defined as

$$Y_i = \begin{cases} 1\ if\ individual\ i\ is\ a\ VCD\ case \\ 0\ otherwise \end{cases}$$

For this section, we are ignoring the cluster membership of participants and combining information across all clusters. Since both the outcome and the exposure defined in the previous section are binary variables, an intuitive simple analysis would be to estimate an odds ratio through the use of a $2\ \times 2$ contingency table. In general, for a 2×2 contingency table that takes the form

	Exposed $(X = 1)$	Unexposed $(X = 0)$
Case $(Y = 1)$	a	b
Non-case $(Y = 0)$	c	d

An odds ratio can be calculated from the contingency table above through the formula $OR = \frac{ad}{bc}$ (20). Woolf's formula for obtaining the variance of the log odds ratio from a 2×2 contingency table is

$$Var(\log(OR)) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

This formula allows one to make inferences about the odds ratio estimate by obtaining a 95% confidence interval for the odds ratio. However, since we do not account for clustering here, the variance calculated is only an approximation, meaning the true variance may be larger and as a result the true 95% confidence interval may be wider.

For this project, we will investigate how the odds ratio changes for different levels of exposure of an individual. To this end, we will assess how the odds ratio changes when an individual is within $[0, 100), [0,200), [0,500), [0,750)$ and $[0,1000)$ metres of a VCD case and also if their illness onsets were less than 7, 14 and 30 days apart, looking at different combinations of time and distance as exposures.

As an illustrative example, we first define a participant as being exposed if they have been within 100 metres of an observed VCD case 7 days before their illness onset. Using this definition of exposure, we can then calculate the number of people that are exposed using the method described in Section 3.2.1 by summing up the number of rows that contain a non-zero element, since that means a participant has been exposed to at least one VCD case. The number of unexposed people is of course the remaining participants which can be calculated by subtracting the number of exposed participants from the total number of participants. We then want to partition the exposed participants into those who ended up being VCD cases and those who ended up being test-negative controls. This is done by filtering the exposed participants to obtain those who are a VCD case, and then the test-negative controls is calculated by finding the difference between VCD cases and total exposed. A similar calculation is done to obtain number of unexposed VCD cases and test-negative controls. Note that the contingency tables will have 385 as the total number of VCD cases as opposed to 392 because we are considering the total number of distinct individuals that were a VCD case and not the number of recorded infections. Following these steps results in the following contingency table shown in Table 2.

	Exposed $(X = 1)$	Unexposed $(X = 0)$
VCD Case $(Y = 1)$	70	315
Test-negative Control $(Y = 0)$	64	5857

Table 2: Contingency table where exposure is being less than 100m of a case within 7 days

From this table, we see that 70 VCD cases had an illness onset within 7 days and had residence less than 100m of another VCD case. Only 64 test-negative controls had illness onset within 7 days and had residence less than 100m of another VCD case. 5857 test-negative controls and 315 VCD cases were each unexposed. Here, we can calculate the odds ratio and the variance of the log odds ratio using the previously stated formulae:

$$OR = \frac{70 \times 5857}{64 \times 315} = 20.34,\ Var(\log(OR)) = \frac{1}{70} + \frac{1}{64} + \frac{1}{315} + \frac{1}{5857} = 0.033$$

Hence, a 95% confidence interval can be obtained by first obtaining a 95% confidence interval for the log-odds ratio and then back-transforming to get the 95% confidence interval for the odds ratio by exponentiating. Following this, we get

$$95\%\ C.I = \left(\exp\left\{\log(20.34) - 1.96 \times \sqrt{0.033}\right\}, \exp\left\{\log(20.34) + 1.96 \times \sqrt{0.033}\right\}\right)$$
$$= (14.24, 29.07)$$

From this, we can see that participants that have resided within 100m of another VCD case and has an illness onset within 7 days of that case has 20.34 times the odds of being a virologically confirmed dengue case than participants who have not been exposed in this way. The 95% confidence interval of (14.24, 29.07) is clearly far above 1, which would provide very strong evidence to reject the null hypothesis that the odds of being a virologically confirmed dengue case is the same amongst the exposed and unexposed groups.

Continuing in this way, we can obtain similar estimates of odds ratios and their respective 95% confidence intervals for the varying combinations of distances and times that define a participant's exposure. The results of these are presented below in Table 3.

Distance	Days	OR	95% CI	Days	OR	95% CI	Days	OR	95% CI
100m	7	20.34	(14.24, 29.07)	14	15.85	(11.76, 21.36)	30	11.26	(8.72, 14.54)
200m	7	9.37	(7.14, 12.29)	14	8.34	(6.58, 10.58)	30	6.83	(5.50, 8.49)
500m	7	5.26	(4.25, 6.52)	14	5.84	(4.71, 7.23)	30	6.72	(5.30, 8.53)
750m	7	4.88	(3.95, 6.03)	14	5.73	(4.53, 7.24)	30	7.07	(5.29, 9.44)
1000m	7	4.38	(3.52, 5.45)	14	5.27	(4.09, 6.80)	30	7.26	(5.15, 10.23)

Table 3: Results from estimating odds ratios for different exposure levels in contingency tables

From Table 3, it is clear that the level of exposure that presents the greatest odds of being a VCD case is when participants reside within 100m of another VCD case and have an illness onset of 7 days or less from that of the VCD case. There is also a negative relationship seen between the odds ratio and distance in which the further away a participant resides from a previously observed VCD case, the lower the odds ratio becomes, and this holds true for when the time constraint is fixed at 7 and 14 days (and is almost true at 30 days).

For example, for participants that have had an illness onset within 7 days of a VCD case, if they reside within 1000m of a VCD case as opposed to 100m, the estimated odds ratio is reduced by nearly 5 times. To this end, it is estimated that participants that reside within 1000m of a VCD and had illness onset within 7 days of that case has 4.38 times the odds of also being a VCD case than participants who were not exposed in this way. The 95% confidence interval of (3.52, 5.45) excludes the null value of 1, therefore providing evidence to reject the null hypothesis of no association between this exposure and being a virologically confirmed dengue case at the 5% level. Similarly, it is estimated that for participants that have had an illness onset within 14 days of a VCD case, if they reside within 1000m of that VCD case as opposed to residing within 100m of the case, the odds ratio is reduced by 3 times. Here, participants that had illness onset within 14 days of a VCD case and resided within 1000m of them had 5.27 times the odds of also being a VCD case than participants who were not exposed in this way. The 95% confidence interval (4.09, 6.80) excludes 1 therefore providing stronger evidence to reject the null hypothesis of no association between this exposure and being a VCD case at the 5% level.

However, interestingly this relationship between distance and the odds ratio changes slightly when the time of illness onset to another VCD case is within 30 days. Table 3 highlights that the odds ratios decrease from 100m to 500m and then steadily increase again once a participant resides within either 750m or 1000m of a VCD case. However, the 95% confidence intervals here also widen suggesting that this observation could be by chance. This may go against what one would expect since, intuitively, we would expect that the further away a participant is from a VCD case, the “safer” they would be so would always expect the odds ratio of being a VCD case to decrease as the distance constraint is relaxed.

Furthermore, the relationship between time and odds ratio appears to be inconsistent for different distances. When a participant resides within either 100m or 200m to a VCD case, Table 3 shows that there is a negative relationship between time and the odds ratio. In this case, as the time of illness onset between a participant and a VCD case increases, the odds ratio of that participant being a VCD case decreases. However, when a participant resides within 500m, 750m, or 1000m to a VCD case, Table 3 shows that as the time of illness onset between a participant and a VCD case increases, the odds ratio of being a VCD case also increases. This latter finding also goes against what one would expect since the greater the difference in time between illness onsets for participants and a VCD case, the less likely they should be to be a VCD case. A comparison of the odds ratios for the different times and distances can be seen in Figure 4 below.

Overall, using contingency tables have been an insightful first step in identifying how spatial and temporal information can be utilised to estimate various odds ratios of participants being a virologically confirmed dengue case. However, this method may be too simple to obtain clinically meaningful information as using these aggregated contingency tables do not consider other factors such as the clusters that participants reside in which may provide further useful information. This provides the motivation to use mixed models and generalised estimating equations.

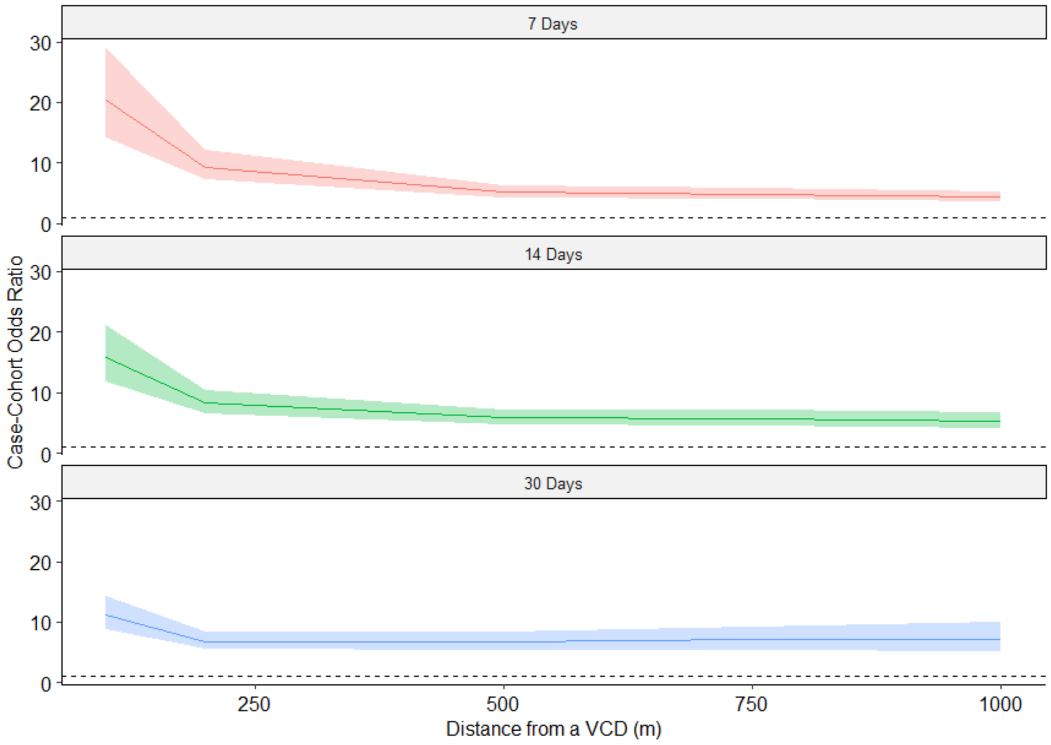

Figure 4: Plots of odds ratio estimates from contingency tables for various times and distances

3.3. Generalised Linear Mixed Models

3.3.1. Introduction

As we are using the same outcome as defined in Section 2.4, if we are to use a regression model to model dengue exposure against binary outcome, a generalised linear model would be the appropriate choice. In particular, a logistic regression model would be the model of choice. However, since we also want to incorporate information provided by the clusters that participants reside in, a hierarchical model would be necessary here. This gives rise to the use of a generalised linear mixed model.

3.3.2. Methods

The general formulation of the generalised linear mixed models used for this analysis is as follows. First, we define:

$Y_{ij} = \begin{cases} 1 \; if \; participant \; i \; is \; a \; VCD \; case \; in \; cluster \; j \\ 0 \; otherwise \end{cases}$ for $j = 1, \dots, 24$

$X_{1ij} = \begin{cases} 1 \; if \; participant \; i \; is \; in \; cluster \; j \; and \; within \; [d_1, d_2) \; metres \; of \; a \; VCD \; case \; within \; t \; days \\ 0 \; otherwise \end{cases}$

$X_{2ij} = \begin{cases} 1 \; if \; individual \; i \; is \; in \; intervention \; cluster \; j \\ 0 \; otherwise \end{cases}$

$\pi_{ij} \; to \; be \; the \; probability \; that \; individual \; i \; residing \; in \; cluster \; j \; becomes \; a \; VCD \; case$

$b_j \; to \; be \; the \; random \; intercept \; corresponding \; to \; cluster \; j$

$\beta_0, \beta_1, \beta_2, \beta_3 \; are \; parameters \; to \; be \; estimated$

Then, the generalised linear mixed model (GLMM) is given by:

$$Y_{ij} | \, b_j \sim Bernoulli(\pi_{ij})$$

$$logit(\pi_{ij}) = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{2ij} + \beta_3 X_{1ij} * X_{2ij} + b_j$$

$$b_j \sim N(0, \sigma_0^2)$$

Here, the fixed effects in the GLMM are exposure status, X_{1ij}, the intervention cluster status, X_{2ij}, and an interaction term between exposure and intervention. The random effect is the random intercept b_j. A random intercept was included because the clusters are heterogeneous which is seen as they vary greatly in terms of the number of participants enrolled in each cluster despite the clusters being arranged so that they are roughly equal in geographical size. This may introduce some variability within clusters, so my introducing a random intercept into the logistic regression model, this will target the within-cluster variability.

Similar to Section 3.2.1, this model was fitted using different levels of exposure which are as defined in that section. Normally, one may wish to check the necessity of a generalised linear mixed model over a logistic regression model with no fixed effects by conducting a hypothesis test on the variance of the random intercept b_j (21). To this end, the null hypothesis is $H_0\colon \sigma_0^2 = 0 \text{ vs } H_1\colon \sigma_0^2 > 0$**.** Since the null hypothesis is on the boundary of the parameter space of the variance, under H_0, the Z-statistic for this hypothesis test follows a positive distribution 50% of the time and will equal zero for the other 50% of the time (21).

Therefore, we would refer to a mixed $\chi^2_{0,1}$ distribution as the null distribution make to inference (21). The test statistic here would be $-2 \times$ log-likelihood of the GLMM and logistic regression model at their respective maximum likelihood estimates. However, since there are 15 separate GLMMs being fitted here, repeating this process to assess the need for a GLMM would certainly increase the probability of type I error due to repeated hypothesis tests. Furthermore, this was not checked as it was decided a priori that GLMMs would be used and that a GLMM would be fit for each definition of exposure for a matter of consistency.

The main difference between GLMMs and standard linear mixed models lies in the interpretation of the two sets of models. For normal linear mixed models, the fixed effect parameters have interpretations which are interchangeably marginal and conditional due to the link function for a linear mixed model being the identity link. However, since the GLMM used in this setting uses the logit link which is non-linear, taking the expectation of the linear predictor does not result in the random effect being cancelled out. Hence, there is no marginal interpretation of the GLMM, resulting in only an interpretation that is conditional on the random effects used in the model which is why the family of GLMMs are also known as subject-specific models.

3.3.3 Results

Similar to the results found in Table 3 from analysing the contingency tables, there is a clear overall negative relationship between the distances at which participants reside from a VCD case and odds ratios relating to exposure, now adjusting for intervention and the interaction term. For example, where exposure is defined as a participant residing within 100m to a VCD case with an illness onset within 7 days of that case, it is estimated that exposed participants have 12.95 (95% CI: 8.68, 19.33) times the odds of being a VCD case than unexposed participants, adjusting for intervention status and interaction term, for participants in a given cluster. Whereas when exposure is defined as a participant residing within 1000m of a VCD case and having an illness onset within 7 days of that case, it is exposed participants have 4.32 (95% CI: 3.38, 5.51) times the odds of being a VCD case than unexposed participants, adjusting for intervention and interaction terms, for participants in a given cluster. Similar findings are observed when illness onset to a VCD case is within 14 days. When illness onset to a case is within 30 days, the estimated odds ratios for exposure, adjusted for intervention and interaction increase at 500m, but since the 95% CI get wider, this may be due to chance. Furthermore, we see that the effect of the intervention, adjusted for exposure and interaction becomes less effective as the distance of a participant to a VCD case increases. When exposure is defined as residing within 100m to a VCD case and having illness onset within 7 days of that case, it is estimated that participants in intervention clusters have 74% (95% CI: 59%, 84%) reduced odds of being a VCD case than participants in untreated clusters, adjusting for exposure and interaction terms, for participants in a given cluster. This protective effect is estimated to reduce to 50% (95% CI: 14%, 71%) when exposure is where participants are instead within 1000m of a VCD case and have illness onset within 7 days of that case. Most notably, there is strong evidence of an association between exposure and being a VCD case, adjusted for intervention and interaction terms across all spatio-temporal combinations used here as highlighted in Table 4 since all of the p-values are <0.01 for the exposure variable. Similarly, there is strong evidence of an association between intervention and being a VCD case, adjusted for exposure and interaction terms across all spatio-temporal combinations up until where participants reside within 1000m of a VCD case and have illness onset within 30 days of that case in which there is only weak evidence (p=0.06). There is only evidence of an interaction between exposure and intervention status when a participant resides within 1000m of a VCD case, and this holds true across all illness onset times. For example, when a participant resides within 1000m of a VCD and has illness onset within 7 days, it is estimated that the odds of an exposed individual becoming a VCD case is reduced by 50% (95% CI: 14%, 71%) if they are in an intervention cluster than if they were exposed in an untreated cluster, for participants in a given cluster, with this showing strong evidence (p=0.01) of an interaction between exposure and intervention statuses.

Distance	Days	Variable	OR (95% CI)	p-value
100m	7	Exposure	12.95 (8.68, 19.33)	<0.01
		Intervention	0.26 (0.16, 0.41)	<0.01
		Exposure * Intervention	1.52 (0.52, 4.43)	0.44
200m	7	Exposure	7.17 (5.29, 9.73)	<0.01
		Intervention	0.28 (0.17, 0.46)	<0.01
		Exposure * Intervention	0.91 (0.39, 2.16)	0.84
500m	7	Exposure	4.34 (3.42, 5.51)	<0.01
		Intervention	0.32 (0.19, 0.54)	<0.01
		Exposure * Intervention	0.67 (0.35, 1.27)	0.22
750m	7	Exposure	4.40 (3.47, 5.58)	<0.01
		Intervention	0.34 (0.19, 0.58)	<0.01
		Exposure * Intervention	0.63 (0.36, 1.09)	0.10
1000m	7	Exposure	4.32 (3.38, 5.51)	<0.01
		Intervention	0.37 (0.21, 0.65)	<0.01
		Exposure * Intervention	0.50 (0.29, 0.86)	0.01
100m	14	Exposure	10.77 (7.74, 15.00)	<0.01

		Intervention	0.28 (0.18, 0.45)	<0.01
		Exposure * Intervention	1.06 (0.40, 2.82)	0.91
200m	14	Exposure	6.82 (5.22, 8.91)	<0.01
		Intervention	0.31 (0.19, 0.52)	<0.01
		Exposure * Intervention	0.60 (0.28, 1.33)	0.21
500m	14	Exposure	4.91 (3.86, 6.23)	<0.01
		Intervention	0.34 (0.19, 0.58)	<0.01
		Exposure * Intervention	0.77 (0.45, 1.34)	0.36
750m	14	Exposure	5.05 (3.91, 6.53)	<0.01
		Intervention	0.36 (0.20, 0.66)	<0.01
		Exposure * Intervention	0.65 (0.38, 1.12)	0.12
1000m	14	Exposure	4.77 (3.63, 6.27)	<0.01
		Intervention	0.39 (0.21, 0.73)	<0.01
		Exposure * Intervention	0.56 (0.32, 0.98)	0.04
100m	30	Exposure	7.90 (5.95, 10.48)	<0.01
		Intervention	0.30 (0.19, 0.48)	<0.01
		Exposure * Intervention	1.06 (0.46, 2.44)	0.89
200m	30	Exposure	5.63 (4.41, 7.19)	<0.01
		Intervention	0.34 (0.21, 0.57)	<0.01
		Exposure * Intervention	0.61 (0.31, 1.19)	0.14
500m	30	Exposure	5.47 (4.21, 7.10)	<0.01
		Intervention	0.37 (0.21, 0.68)	<0.01
		Exposure * Intervention	0.71 (0.41, 1.22)	0.21
750m	30	Exposure	5.52 (4.09, 7.46)	<0.01
		Intervention	0.38 (0.20, 0.74)	<0.01
		Exposure * Intervention	0.64 (0.35, 1.17)	0.14
1000m	30	Exposure	5.71 (4.05, 8.05)	<0.01
		Intervention	0.51 (0.25, 1.02)	0.06
		Exposure * Intervention	0.43 (0.23, 0.80)	<0.01

Table 4: Results from estimating odds ratios using generalised linear mixed models

One difference observed between the estimated odds ratios relating to exposure obtained using generalised linear mixed models and the estimated odds ratios obtained using the contingency tables is that the estimated odds ratios from the generalised linear mixed models are all smaller in magnitude compared to the odds ratios from the contingency tables. This may be because once the participants’ intervention status and the potential interaction between exposure and intervention is taken into account, it explains some of the relationship between this spatial-temporal exposure and the odds of being a VCD case, therefore reducing the exposure effect. In addition, one similarity between these two methods of estimation is that for both methods, the relationship between difference in time of illness between a participant and a VCD case and the odds ratio relating to exposure is negative for shorter distances (100m and 200m) of proximity to a VCD case and positive for longer distances of proximity to a VCD case.

3.3.4. Discussion

Overall, using generalised linear mixed models have shown strong evidence of an association between the odds of being a VCD case and both exposure and intervention respectively. A clear relationship is also observed between the proximity of a participant to a VCD case and the resulting odds ratios relating to exposure, adjusting for intervention and the interaction, and the odds ratios relating to intervention, adjusting for exposure and the interaction, given participants are in the same cluster. However, one downside of using these models is that whilst they take the cluster of participants into consideration, due to the nature of generalised linear mixed models being subject-specific models, we are unable to make any direct marginal interpretations from these results and marginal interpretations may be more useful from a public health standpoint. In order to obtain marginal interpretations, we can fit the population-averaged models through the use of generalised estimating equations.

3.4. Generalised Estimating Equations

3.4.1. Introduction

As mentioned in 3.3.4, the need for generalised estimating equations lie in the fact that the interpretations of parameters used in this model are marginal and can therefore be used to make inferences about the population as opposed to specific clusters as provided by the generalised linear mixed model. Generalised estimating equations only require correct specification of the univariate marginal distributions provided we are willing to adopt ‘working’ assumptions about the association structure (21). These models estimate parameters associated with the expected value of an individual’s vector of binary responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations; it combines estimating equations for the regression parameters with moment-based estimation for the correlation parameters entering the working assumptions (21). A benefit of using generalised estimating equations is that provided the model for the mean is specified correctly, the resulting estimates of the parameters are asymptotically consistent and normally distributed. Another benefit of using generalised estimating equations is that they do not require the correct specification of the working correlation structure because point estimates and empirically-corrected standard errors remain asymptotically correct (21). However, there may be some loss in efficiency if the working correlation structure differs strongly from the true underlying structure (21).

3.4.2. Methods

The model used in generalised estimating equations is quite similar to that of the generalised linear mixed model, except it does not include any random effects. Similar to previous sections, models were fitted for various spatial and temporal combinations that defined exposure. The general format of the generalised estimating equations model is as follows:

Define $Y_{ij}, X_{1ij}, X_{2ij}, \pi_{ij}, \beta_0, \beta_1, \beta_2, \beta_3$ as in Section 3.3.2, we have that

$$Y_{ij} \sim Bernoulli(\pi_{ij})$$

$$logit(\pi_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{1ij} * X_{2ij}$$

Furthermore, the independence working correlation structure was used to fit these models. This is because independence and exchangeable working correlation structures can be used in almost all applications such as clustered, longitudinal or correlated data, whereas the AR(1) and unstructured working correlation structures are less relevant to our clustered data. Although selecting either of the latter two structures would not be wrong, the independence working correlation structure was chosen over the exchangeable working correlation structure due to its simplicity. This is because the independence working correlation structure uses the identity matrix whereas the exchangeable working structure does not. By using the independence working correlation structure, it means we assume that participants within each cluster are independent of each other as a working assumption (but use robust standard errors to allow for unknown levels of within cluster correlation).

3.4.3. Results

The results from running several generalised estimating equation models are found in Table 5 below.

Distance	Days	Variable	OR (95% CI)	p-value
100m	7	Exposure	13.72 (6.52, 28.86)	<0.01
		Intervention	0.25 (0.17, 0.38)	<0.01
		Exposure * Intervention	1.76 (0.40, 7.77)	0.46
200m	7	Exposure	6.93 (4.59, 10.47)	<0.01
		Intervention	0.26 (0.18, 0.38)	<0.01
		Exposure * Intervention	1.00 (0.27, 3.65)	1.00
500m	7	Exposure	4.19 (2.90, 6.04)	<0.01

		Intervention	0.30 (0.21, 0.45)	<0.01
		Exposure * Intervention	0.66 (0.28, 1.55)	0.34
750m	7	Exposure	4.19 (3.12, 5.63)	<0.01
		Intervention	0.32 (0.20, 0.51)	<0.01
		Exposure * Intervention	0.61 (0.33, 1.14)	0.12
1000m	7	Exposure	4.17 (3.14, 5.54)	<0.01
		Intervention	0.36 (0.22, 0.59)	<0.01
		Exposure * Intervention	0.47 (0.27, 0.82)	0.01
100m	14	Exposure	11.01 (5.93, 20.45)	<0.01
		Intervention	0.27 (0.18, 0.39)	<0.01
		Exposure * Intervention	1.24 (0.33, 4.65)	0.75
200m	14	Exposure	6.46 (4.57, 9.14)	<0.01
		Intervention	0.29 (0.20, 0.44)	<0.01
		Exposure * Intervention	0.67 (0.25, 1.78)	0.42
500m	14	Exposure	4.65 (3.40, 6.37)	<0.01
		Intervention	0.31 (0.19, 0.51)	<0.01
		Exposure * Intervention	0.74 (0.44, 1.27)	0.27
750m	14	Exposure	4.78 (3.48, 6.57)	<0.01
		Intervention	0.34 (0.20, 0.57)	<0.01
		Exposure * Intervention	0.62 (0.36, 1.07)	0.08
1000m	14	Exposure	4.57 (3.34, 6.24)	<0.01
		Intervention	0.37 (0.21, 0.65)	<0.01
		Exposure * Intervention	0.52 (0.30, 0.90)	0.02
100m	30	Exposure	7.87 (5.05, 12.27)	<0.01
		Intervention	0.28 (0.19, 0.40)	<0.01
		Exposure * Intervention	1.14 (0.28, 4.57)	0.86
200m	30	Exposure	5.33 (3.96, 7.18)	<0.01
		Intervention	0.32 (0.22, 0.47)	<0.01
		Exposure * Intervention	0.63 (0.28, 1.41)	0.26
500m	30	Exposure	5.21 (4.01, 6.78)	<0.01
		Intervention	0.35 (0.22, 0.56)	<0.01
		Exposure * Intervention	0.67 (0.39, 1.13)	0.13
750m	30	Exposure	5.27 (3.82, 7.23)	<0.01
		Intervention	0.37 (0.22, 0.61)	<0.01
		Exposure * Intervention	0.60 (0.34, 1.08)	0.09
1000m	30	Exposure	5.52 (3.74, 8.15)	<0.01
		Intervention	0.49 (0.28, 0.86)	0.01
		Exposure * Intervention	0.40 (0.21, 0.75)	<0.01

Table 5: Results from estimating odds ratios using generalised estimating equations

Here, the main difference in the interpretations of these odds ratios is that they are marginal, meaning they apply to the whole population as opposed to specific clusters, which has its use for implementing public health policies and are thus comparable to the earlier contingency table estimates. For example, when exposure is defined as participants residing within 100m of a VCD case and has illness onset within 7 days of that case, it is estimated that exposed participants have 13.72 (95% CI: 6.52, 28.86) time the odds of being a VCD case than unexposed participants, adjusting for intervention cluster and the interaction term. In addition, when exposure is defined in the same way, it is estimated that participants in intervention clusters have 75% (95% CI: 62%, 83%) lower odds of being a VCD case than participants in untreated clusters, adjusting for exposure status and the interaction term. The results from the generalised estimating equations are fairly similar to the results using generalised linear mixed models as there is also a negative relationship between a participant's proximity to a VCD case and the odds ratio for exposure, adjusting for intervention and interaction terms when illness onset from a VCD case is within 7 days. When illness onset is within 14 and 30 days, the results here show a slight increase in this odds ratio after 500m, but similar to Section 3.3.3, the 95% CIs widen as distance increases meaning this increase could be due to chance. Another similarity is that both models show strong evidence of association between odds of being a VCD case and exposure, adjusted for intervention and interaction, and the odds of being a VCD case and intervention, adjusted for exposure and interaction as all the p-values relating to these variables are $p \leq 0.01$. Analogous to Section 3.3.3, we see that there is only evidence of an interaction between exposure status and intervention status when proximity of a participant to a VCD case is within 1000m, across all illness onset times. Furthermore, similar to both the contingency table method and generalised linear mixed model method, the relationship between a participant's difference in time of illness onset to a VCD case and the estimated odds ratio for exposure is only decreasing for shorter proximities and increases for larger proximities.

3.5. Discussion

In conclusion, all three methods used in this section provided insight into how different levels of exposure defined by distance and time changes the odds of a participant becoming a VCD case, obtaining both conditional and marginal interpretations. The latter two methods also explored the association between intervention status and odds of being a VCD case, adjusting for exposure and the interaction between exposure and intervention status.

Comparing the odds ratios relating to exposure status across different times of illness onset, the estimated odds ratios obtained from the contingency tables are consistently larger than the estimated odds ratios obtained from generalised estimating equations and generalised linear mixed models. This may be due to the contingency tables not taking into account the intervention status of participants or the interaction between intervention and exposure, therefore overestimating the exposure effect. The estimates obtained from the generalised linear mixed models and generalised estimating equations appear to be roughly similar, with both models agreeing with regards to making statistical inferences about the variables. The only difference is that when exposure is defined as a participant residing within 1000m of a VCD case and having illness onset within 30 days of that case, the GLMM provides only weak evidence (p=0.06) of an association between intervention and risk of dengue, adjusting for exposure and interaction. Whereas in the GEE model provides strong evidence (p=0.01) of an association between intervention and risk of dengue, adjusting for exposure and interaction.

Some ways in which these estimates could be improved is if there could be a way to incorporate the people who were infected with DENV twice as their second infection was excluded from the analysis when their residence when they were first infected may have been different to their residence at the second location which may have provided some extra information. Furthermore, the generalised linear mixed models and generalised estimating equations could be improved if the data contained some additional information on participants that could be used as fixed effects and refine the mean function as an assumption from the generalised estimating equations is that the mean function is correctly specified which it may not be in reality.

4. Spatial Dependence and Tau

4.1. Introduction

4.1.1. Introduction to Spatial Data

The methods used throughout Section 3 to identify how different levels of ‘exposure’ based on a participant’s distance and difference in time of illness onset relating to a VCD case are associated with their odds of also being a VCD case can be seen as an introduction to the methods more traditionally used to assess the existence of spatio-temporal clustering of a disease. To this end, spatial data are defined as observations in which labels have been added to show where observations were “collected” (22). In this context, the AWED trial data are clearly spatial data as participants provided information on their location of primary residence which was then converted into latitude and longitude.

In Geographical Information Systems (GIS), vector data refers to spatial data that consists of the following spatial types: Points which are single point locations such as a house; lines which are a collection of ordered points that are connected such as a road; polygons which are a series of connected points where the first and last point are in the same location such as a lake or administrative area (22). Note that polygons can also contain holes. For cluster randomised trials that include geographical information, the vector data model is most commonly used to describe spatial representation (22). In the context of our data, the polygons are represented by the clusters and the points are the location of a participant’s primary residence.

Vector data which has points that are the location of a person’s home, place of likely disease acquisition or other relevant measure of spatial location are commonly referred to in the literature as point pattern data (23). Point pattern data which also include the time of occurrence of the event of interest are known as spatio-temporal point process data (24). Hence, the AWED trial data can be referred to as spatio-temporal point process data since the date of illness onset for each participant was also recorded during the clinical visit. When faced with point pattern or point process data, it is usually of primary interest to investigate the population dynamics of the infectious disease using the spatial or temporal scales over which cases are more likely to be found (23). Now, we introduce the concept of spatial clustering.

4.1.2 Spatial Clustering and Measures

Unlike the usual epidemiological definition for a cluster, in the context of spatial statistics clustering is when points tend to appear closer than each other than would be expected if they were completely spatially randomly distributed (23). Measures of clustering are broadly divided into two types: local clustering statistics which measure the tendency of events to occur around a particular point in space, and global clustering statistics which measure the tendency of events to cluster in space in general (23). Both measures use point pattern data as input and due to the nature of our data being points (cases) over a broad space, we will be focusing on global clustering statistics.

The location of cases in relation to their infectors define a spatial transmission kernel for an infectious disease which is a probability distribution of distances between the locations of an infector and infectee (23). The shape of a spatial transmission kernel depends on factors such as host behaviour, mode of transmission and environment (23). In the case of dengue, since *Aedes Aegypti* mosquitoes breed in both indoor and outdoor water containers, the environment and mode of transmission have a more important role than host behaviour in the defining of a spatial transmission kernel. The way in which host, mode of transmission and environment interact to determine a transmission kernel can be complex and hard to measure (23).

An example of a transmission kernel used in practice is seen in (24) in which it was the central feature of the model used define the conditional rate of transmission of foot and mouth disease in cows between farms in Cumbria, England.

Although it is intuitively natural to specify a model for a spatio-temporal point process through its conditional intensity at location x at time t given the history of the process up to time t, this typically results in an analytically intractable likelihood (24). Furthermore, even if clustering in where people live is taken into account, the actual distribution of related cases that we see in the population represents the results of multiple generations of transmission (23) and population density. In many situations, the observed spatial clustering of cases after multiple generations of transmission may be of greater public health importance than the spatial transmission kernel itself (23), showing the need for other measures of global clustering which do not rely on spatial transmission kernels.

There are a number of measures and methods that can be used to obtain global clustering statistics or to model spatial clustering by either utilising the exact locations of points or by aggregating them into grid cells (23). In terms of modelling spatial dependence through the use of regression models, when spatial correlation (or clustering) is present then nearby observations are related to one another and it typically can be identified through the presence of spatially correlated residuals in these models (22). This clearly violates one of the main assumptions of the linear regression model being the independence of error terms. Since the mechanism through which spatial correlation is rarely observable, spatial models use latent variables and structures to incorporate spatial effects (22). The standard approach is to use a spatial weights matrix which is a mathematical representation of the spatial structure of the data or by redefining the covariance matrix of the error term in a linear regression model which gives rise to the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models; further details can be found in (22). Further global clustering statistics include Moran's I which is used to assess the existence of spatial autocorrelation in spatial data (22, 23). It is an extension of Pearson's product-moment correlation into two dimensions by considering the strength of association and spatial lag over which it is present (22). Another popular statistic known as the K-function is defined as

$$K(d) = \lambda^{-1} E[N(d)]$$

Where λ is the spatial intensity of points in the study area (calculated as study area size divided by number of points) and $E[N(d)]$ is the expected number of points within distance, d, of a point (23). In the special case of a homogeneous Poisson process, the value of the K-function is πd^2 and this theoretical value can be used to compare with an estimated K-function of a point process to identify the presence of clustering (23).

Although techniques such as K-function and Moran's I are commonly used and prove to be useful, for a global clustering statistic to be useful in the field of infectious disease epidemiology, it would ideally be easily interpretable in terms of disease risk, be comparable across different settings and distinguish spatial variation due to the transmission process or due to clustering in the underlying population (23). These two techniques and a number of others do not necessarily satisfy these criteria. To this end, (23) proposes a natural measure of the clustering of disease risk known as the τ- statistic.

4.1.3. τ-statistic

As highlighted in the previous section, epidemiologists characterise populations or exposures leading to elevated risk of a health outcome through the use of a relative risk measure or approximation (23). These measures include odds ratios used throughout Section 3, risk ratios, incidence ratios, etc. Here, we introduce the τ-statistic which is a measure of relative risk of someone at a particular spatial distance from a case also being a case, versus the risk of anyone in the population being a case (23). It is given by

$$\tau(d_1 . d_2) = \frac{\lambda(d_1, d_2)}{\lambda}$$

Where $\lambda(d_1, d_2)$ is the expected incidence rate of someone in the distance range (d_1, d_2) of a case and λ is the average incidence across the whole population.

If the study population is known and the incidence rate is constant over some period of time, then τ can be trivially estimated as

$$\hat{\tau}(d_1, d_2) = \frac{\hat{\pi}(d_1, d_2)}{\hat{\pi}(0, \infty)}$$

where $\hat{\pi}(d_1, d_2)$ is the estimated probability that a case occurs within distance range (d_1, d_2) of another case. Note that test-negative controls are included in this estimate. For this case where the population is known, the interpretation of this statistic is quite simple. For example, suppose the study population in the AWED trial was known, then $\tau(0, 500m) = 2$ would mean that VCD cases are twice as common within 500 metres of another VCD case than they are in the whole population. Another benefit of using this statistic is that unlike the K-statistic, τ does not require the need for edge corrections which are used to account for the usually arbitrary location of study area borders; a discussion of this can be found in the supplementary text of (23).

However, in most cases including the case for the AWED trial, the underlying distribution of the population at risk is unknown. In these cases, we have to adapt the τ-statistic to use information on the infecting pathogen such as serotype to distinguish between pairs of cases that are consistent with being in the same transmission chain and pairs that are inconsistent with coming from the same chain (23). Virologically confirmed dengue cases that have the same serotype are called homotypic cases whereas virologically confirmed dengue cases that have different serotypes are called heterotypic cases (23). Since the *Aedes Aegypti* mosquito would only carry one serotype of DENV, it is only possible for pairs of homotypic cases to come from the same transmission chain, meaning pairs of heterotypic cases must come from different transmission chains. If $\pi(d_1, d_2)$ is approximated by the proportion of cases within a spatial region that are homotypic, this will bias the spatial estimate towards the null value of 1; this is discussed in the supplementary text of (23). However, there does exist a valid estimator of the τ-statistic and it is given by

$$\hat{\tau}(d_1, d_2) = \frac{\hat{\theta}(d_1, d_2)}{\hat{\theta}(0, \infty)} \qquad (1)$$

Where $\hat{\theta}(d_1, d_2)$ is an estimator of the odds that a case within distance (d_1, d_2) of a case is homotypic (23).

As the odds estimator above focuses on homotypic cases, it assumes that the probability of a pair of cases occurring within (d_1, d_2) being heterotypic is the same as the probability of a heterotypic pair of cases occurring anywhere in the population (23). However, even if we believe that this assumption is not true, this estimator is still valid; although it does mean that its interpretation is slightly altered (23). Now, the interpretation of $\hat{\tau}$ is that it is a measure of the clustering in potentially related cases over and above the clustering of unrelated cases due to secular factors such as environmental conditions or population density (23, 25). In the context of our AWED trial data, the numerator of $\hat{\tau}$ identifies, among those enrolled within a particular space-time window, the number of homotypic dengue pairs relative to the number of pairs of enrolled individuals who are assumed to be non-transmission related (25). This group of individuals who are assumed to be non-transmission related include both the test-negative controls, heterotypic cases or homotypic cases that occur beyond a pre-specified time. It should also be noted that although $\hat{\tau}(d_1, d_2)$ is used loosely as an odds ratio here, there are subtle differences between this estimator and the standard epidemiological odds ratio parameter that do not make them equal (25).

4.2. Methods

In order to identify the existence of any spatial dependence present in the AWED trial data, we use the $\hat{\tau}$-statistic as our estimator here. Since this estimator relies on the differentiation between homotypic and heterotypic cases, it was necessary to exclude from this analysis the 67 VCD participants whose DENV serotype was recorded as unknown. A map showing the virologically confirmed dengue cases in the Yogyakarta study area partitioned by serotype (including unknown) is shown below in Figure 5.

Here, $\hat{\tau}$ was calculated at arm level and then compared to assess if introducing the *wMel*-infected *Aedes Aegypti* mosquitoes disrupts any existing spatial dependence. Once the participants with an unknown dengue serotype were filtered out of the data set, the remaining participants were partitioned into two data sets by intervention status. As well as comparing $\hat{\tau}$ between the two arms, we compared $\hat{\tau}$ within the arms for different distances and timeframes. To this end, $\hat{\tau}(0, 100m), \hat{\tau}(0, 200m), \hat{\tau}(0, 500m), \hat{\tau}(0, 750m)$, and $\hat{\tau}(0, 1km)$ were calculated each for the cases where illness onset between pairs of participants occurred within 7 days, 14 days and finally 30 days of each other.

To calculate $\hat{\tau}(d_1, d_2)$, it required the use of a permutation function written by Suzanne Dufault of University of California, Berkeley which goes as follows:

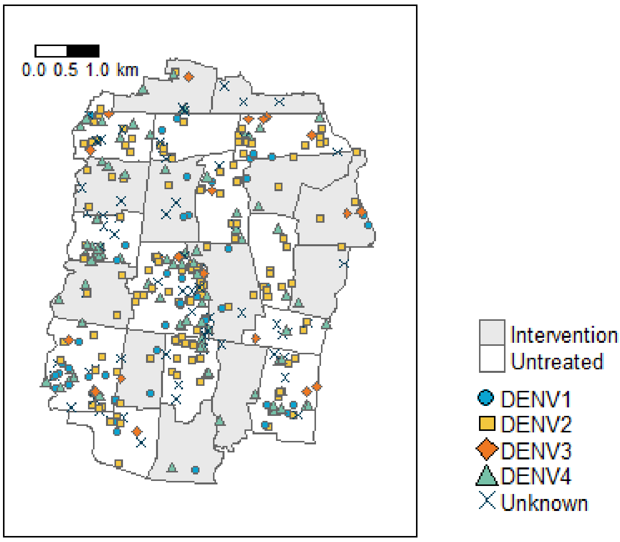

Figure 5: A map of the Yogyakarta study area including the VCD cases by serotype status

Recall (1), the equation that defines the estimate for $\hat{\tau}(d_1, d_2)$ where the population is unknown. We define:

$$z_{ij} = \begin{cases} 1 \, if \; indivduals \; i \; and \; j \; are \; potentially \; transmission \; related \\ 0 \, if \; individuals \; i \; and \; j \; are \; not \; potentially \; transmission \; related \end{cases}$$

$$\hat{\theta}(d_1, d_2) = \frac{\Sigma_i \Sigma_j \mathbb{1}(z_{ij}=1,\; d_1 < d_{ij} < d_2)}{\Sigma_i \Sigma_j \mathbb{1}(z_{ij}=0,\; d_1 < d_{ij} < d_2)}$$

Hence, we are calculating the number of concordant pairs over the number of discordant pairs. To estimate this, it requires four different quantities being:

- No. of transmission related homotypic pairs in a space-time window (a)
- No. of non-transmission related pairs in a space-time window (b)
- No. of transmission related homotypic pairs in a time window over whole study area (c)
- No. of non-transmission related pairs in a time window over whole study area (d)

Then we have that $\hat{\tau} = (a/b)/(c/d)$. These four quantities can be calculated by defining the following matrices:

$\mathbb{A}$ be a $n \times n$ matrix of pairwise absolute differences in enrolment time (for n individuals)
$\mathbb{B}$ be a $n \times n$ matrix of pairwise distances in meters
$\mathbb{C}$ be a $n \times n$ matrix of homotypic pairs (z_{ij})

Then, for a given time t and distance range (d_1, d_2), we can transform $\mathbb{A}$ and $\mathbb{B}$ into binary matrices using the conditions: $\mathbb{A} \leq t$ and $d_1 < \mathbb{B} < d_2$ and we arrive at the four quantities by the following matrix multiplications:

$$a = \sum \mathbb{A} \times \mathbb{B} \times \mathbb{C}$$

$$b = \sum \mathbb{A} \times \mathbb{B} \times (1 - \mathbb{C})$$

$$c = \sum \mathbb{A} \times \mathbb{C}$$

$$d = \sum \mathbb{A} \times (1 - \mathbb{C})$$

Note that pairs including test-negative controls cannot contribute to the numerator of θ.

Once obtaining our estimate of $\hat{\tau}(d_1, d_2)$, our next interest is to have some form of uncertainty present in order to tell if spatial dependence actually exists. First, we obtain a permutation-based null distribution for $\hat{\tau}(d_1, d_2)$. This is done by permuting the participant location data and then calculating $\hat{\tau}(d_1, d_2)$ using these permuted locations and this was repeated 1000 times. A 95% confidence interval was then obtained from this permutation-based null distribution by taking the values of $\hat{\tau}(d_1, d_2)$ at the 2.5th and 97.5th percentiles of the distribution. The null hypothesis here would be that there is no spatial dependence in the transmission of dengue virus versus the alternative hypothesis that there exists some spatial dependence in the transmission of dengue virus.

4.3. Results

4.3.1. Illness onset within 7 Days

Time (Days)	Distance (d_1, d_2) (m)	Treatment	$\hat{\tau}(d_1, d_2)$
7	(0, 100)	Intervention	6.64
7	(0, 200)	Intervention	2.42
7	(0, 500)	Intervention	0.63
7	(0, 750)	Intervention	0.40
7	(0, 1000)	Intervention	0.31

Table 6: Results from tau analysis for intervention arm at 7 days

Time (Days)	Distance (d_1, d_2) (m)	Treatment	$\hat{\tau}(d_1, d_2)$
7	(0,100)	Untreated	5.58
7	(0, 200)	Untreated	3.27
7	(0, 500)	Untreated	1.78
7	(0, 750)	Untreated	1.40
7	(0, 1000)	Untreated	1.29

Table 7: Results from tau analysis for untreated arm at 7 days

The results in Tables 6 and 7 above show the reported estimates from running the permutation function over the given distances when difference in time of illness onset for pairs of participants is less than 7 days. Here, it shows that there is a decreasing relationship between $\hat{\tau}(d_1, d_2)$ and the distance between primary residence of participant pairs in both treatment arms as the value of $\hat{\tau}(d_1, d_2)$ decreases as the distance range expands. The largest estimate is for $\hat{\tau}(0{,}100)$ which is observed in the intervention group. This estimates that an individual within the intervention arm that enrolled within 7 days and resides within 100m of a VCD case in the same arm has 6.64 times the odds to be a homotypic pair of cases (and therefore potentially transmission related) than any individual with illness onset occurring within 7 days across all of the intervention areas. Whereas it is estimated that an individual within the untreated arm that enrolled within 7 days and resides within 100m of a VCD case in the same arm has 5.58 times the odds to be a homotypic pair of cases (and therefore potentially transmission related) than any individual with illness onset occurring within 7 days across all of the untreated areas. Although the intervention arm saw a larger effect when individuals live within 100m of each other, the estimated tau values in this arm decrease faster than the values seen in the untreated arm. In fact, for individuals that live within 500m and beyond, we see that it is estimated that there is a protective effect. For example, it is estimated that an individual within the intervention arm that enrolled within 7 days and resides within 750m of a VCD case in the same arm has the odds of being a homotypic pair of cases reduced by 36% relative to any individual with illness onset occurring within 7 days across all of the intervention areas. While it is not estimated that the untreated arm experiences any protective effect, the tau estimates do appear to tend to the null value of 1 as the distance range increases which makes sense intuitively.

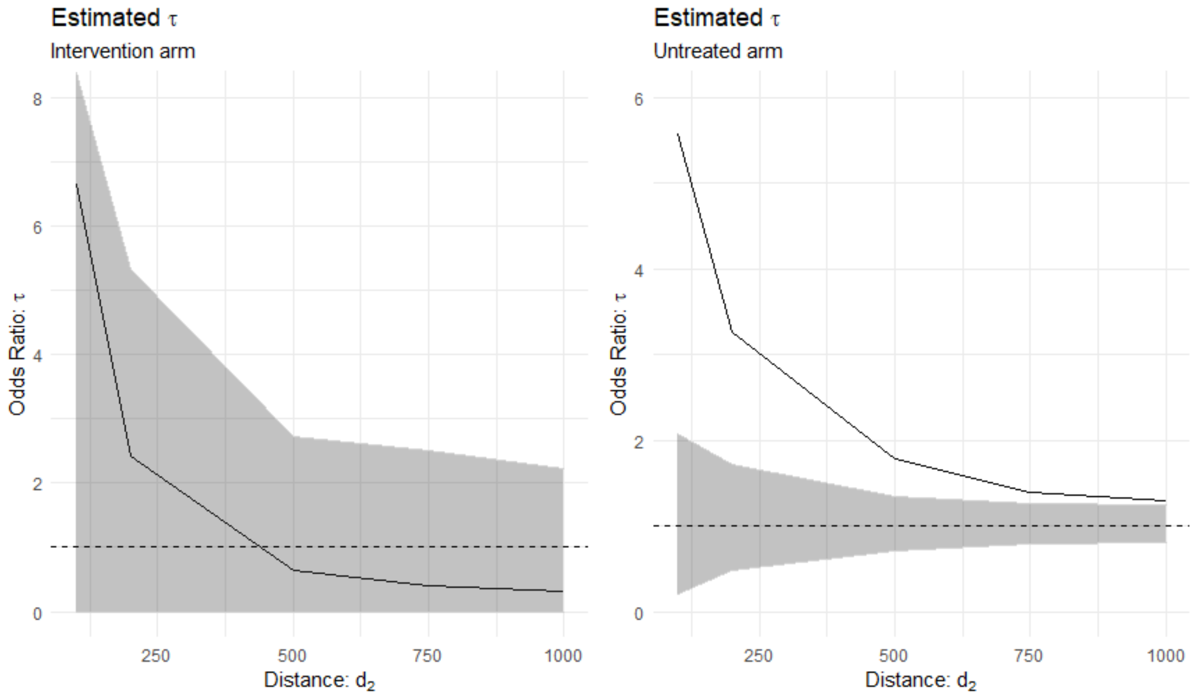

Figure 6: Comparison of tau estimates with 95% confidence intervals for both treatment arms within 7 days of illness onset. Shaded area reflects the permutation null distribution of no spatial clustering.

Furthermore, Figure 6 above highlights that for the intervention arm where illness onset between pairs is within 7 days, there is no evidence to reject the null hypothesis that there is no spatial dependence in the transmission of dengue virus at the 5% level. Here, the observed values of $\hat{\tau}(d_1, d_2)$ fall within the 95% confidence interval of the permutation-based null distribution at each distance range. In contrast, for the untreated arm where illness onset between pairs is within 7 days, there is strong evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue virus. This strong evidence is seen where distance between pairs are within 100m and 200m since $\hat{\tau}(0{,}100m)$ and $\hat{\tau}(0, 200m)$ estimates are far greater than the upper bound of the 95% confidence interval of the permutation-based null distribution. Although the strength of this evidence against the null hypothesis starts to decrease beyond 500m since the tau estimates become closer to the 95% confidence level as the distance range increases; this demonstrates that the strength of spatial dependence decreases as distance between an individual and a VCD case increases.

4.3.2. Illness onset within 14 Days

Time (Days)	Distance (d_1, d_2) (m)	Treatment	$\hat{\tau}(d_1, d_2)$
14	(0, 100)	Intervention	4.79
14	(0, 200)	Intervention	1.69
14	(0, 500)	Intervention	0.88
14	(0, 750)	Intervention	0.55
14	(0, 1000)	Intervention	0.42

Time (Days)	Distance (d_1, d_2) (m)	Treatment	$\hat{\tau}(d_1, d_2)$
14	(0, 100)	Untreated	5.32
14	(0, 200)	Untreated	3.26
14	(0, 500)	Untreated	1.75
14	(0, 750)	Untreated	1.38
14	(0, 1000)	Untreated	1.28

Table 8: Results from tau analysis for intervention arm at 14 days

Table 9: Results from tau analysis for untreated arm at 14 days

Analogous to Section 4.3.1, the results in Tables 8 and 9 show estimated values of $\hat{\tau}(d_1, d_2)$ when the time of illness onset between pairs of individuals are within 14 days of each other. A difference between these results and those found in Tables 6A and 6B is that the estimated tau values for the intervention arm are all consistently lower than the estimated tau values for the untreated arm. For example, now in the case where an individual within the intervention arm enrolled within 14 days and resides within 100m of a VCD case in the same arm, it is estimated that they have 4.79 times the odds to be a homotypic pair of cases (and therefore potentially transmission related) than any individual with illness onset occurring within 14 days across all of the intervention areas. Whereas it is estimated that an individual within the untreated arm that enrolled within 14 days and reside within 100m of a VCD case in the same arm has 5.32 times the odds to be a homotypic pair of cases (and therefore potentially transmission related) than any individual with illness onset occurring within 14 days across all of the untreated areas. It is also clear that, similar to the 7-day case, the negative relationship between estimated tau values and distance range at which pairs of participants reside in holds true in both treatment arms as the estimated tau decreases in both arms as the distance in residence location between pairs increase. Here, we also see a protective effect in the intervention arm beyond 500m since the estimated tau values go below 1 after this distance range.

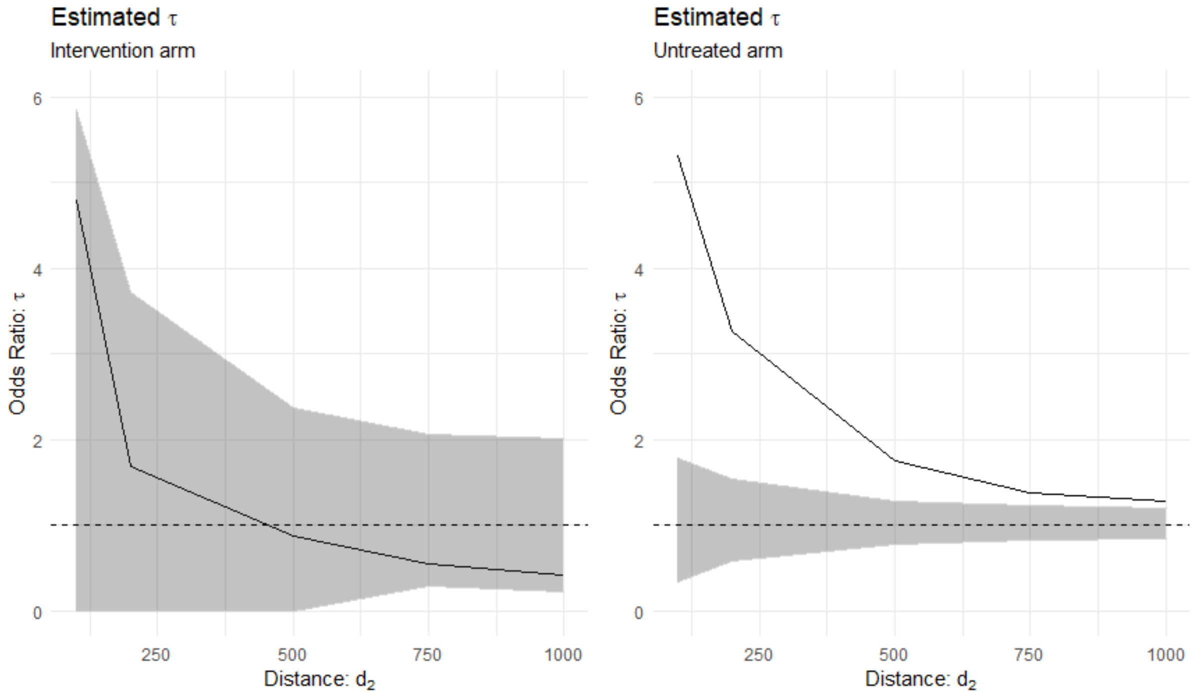

Figure 7: Comparison of tau estimates with 95% confidence intervals between treatment arms within 14 days of illness onset

Also analogous to the case where difference in time in illness onset is 7 days, for the case where difference in time in illness onset between pairs of participants is within 14 days, there is no evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue virus for the intervention arm at the 5% level. This is seen in Figure 7 as the estimated tau values given by the line lies within the 95% confidence interval provided by the permutation-based null distribution. However, there is again strong evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue for the untreated arm at the 5% level. The estimated tau values for shorter proximities are far greater than the null-distribution 95% confidence interval, though this strength of evidence decreases again as proximity increases since the estimated tau approaches the 95% confidence interval for larger distances in residence between pairs.

4.3.3. Illness onset within 30 Days

Time (Days)	Distance (d_1, d_2) (m)	Treatment	$\hat{\tau}(d_1, d_2)$
30	(0, 100)	Intervention	1.98
30	(0, 200)	Intervention	1.34
30	(0, 500)	Intervention	0.87
30	(0, 750)	Intervention	0.99
30	(0, 1000)	Intervention	0.76

Time (Days)	Distance (d_1, d_2) (m)	Treatment	$\hat{\tau}(d_1, d_2)$
30	(0, 100)	Untreated	4.76
30	(0, 200)	Untreated	2.93
30	(0, 500)	Untreated	1.72
30	(0, 750)	Untreated	1.34
30	(0, 1000)	Untreated	1.26

Table 10: Results from tau analysis for intervention arm at 30 days

Table 11: Results from tau analysis for untreated arm at 30 days

The results shown in Tables 10 and 11 show the estimated tau values in each arm where the difference in illness onset between pairs of participants are within 30 days. The value of $\hat{\tau}(0{,}100m)$ in the intervention arm is notably smaller than the cases where the difference in illness onset between participant pairs are shorter. To this end, it is estimated that an individual within the intervention arm that enrolled within 30 days and resides within 100m of a VCD case in the same arm has just under double the odds to be a homotypic pair of cases than any individual with illness onset occurring within 30 days across all of the intervention areas. However, one difference seen here in the intervention arm is that the relationship between the tau estimates and distance between residences of participants is not monotonic in the 30-day case. Here, we see that $\hat{\tau}(0{,}750m)$ slightly increases to 0.99 which is approximately no change in the odds of being a homotypic pair between an individual that resides within 750m of a VCD case and has illness onset within 30 days of that case, and any individual with illness onset occurring within 30 days across the intervention areas. This is due to there being a larger increase in the number of transmission related homotypic pairs in this space-time window than non-transmission related pairs in the space-time window (i.e. a and b as described in Section 4.2), meaning the ratio will therefore increase. A table containing the full breakdown of $a, b, c,$ and d is included in Appendix 1.

Identical to Sections 4.3.1 and 4.3.2, for the case where difference in time in illness onset between pairs of participants is within 30 days, there is no evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue virus for the intervention arm at the 5% level as all of the tau estimates in this arm lie within the 95% confidence interval obtained from the permutation-based null distribution as seen in Figure 8. Also identical to the previous cases, there is strong evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue for the untreated arm at the 5% level in which this strength of evidence reduces as distance increases as the line given by the estimated tau values are outside the 95% confidence interval although it tends towards the upper limit for larger distances. In addition, there appears to be a negative relationship amongst the untreated arm between time of illness onset between pairs and the estimated tau values across each distance combination. As the time between illness onset between a pair in the untreated arm increases, a slight decrease in the estimated tau is observed for each distance. However, for the intervention arm, this negative relationship is only observed when pairs have distance of residences of less than 100m and 200m. For the larger distances, the opposite effect is observed in which the tau estimates increase as time between illness onset increases.

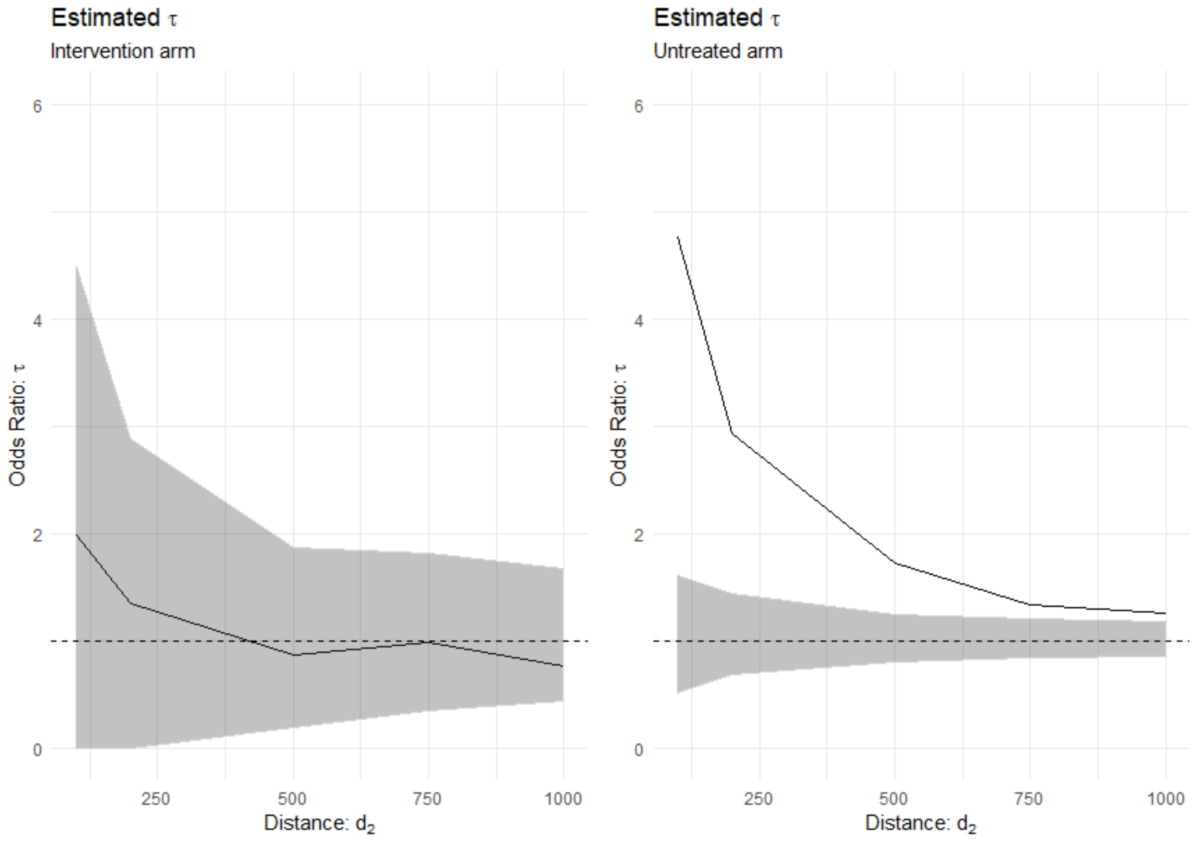

Figure 8: Comparison of tau estimates with 95% confidence intervals between treatment arms within 30 days of illness onset

4.4. Discussion

Overall, using the tau estimates have proven to be very insightful in investigating the effect of the deployment of *wMel*-infected *Aedes Aegypti* mosquitoes in the intervention areas. It is clear that the introduction of these mosquitoes has removed any spatial dependence in the transmission of the dengue virus that existed since there was no evidence to reject this null hypothesis in the intervention arm at any space and time combination. However, it also suggests the existence of general spatial dependence in the transmission of dengue since the untreated arms showed evidence to reject this null hypothesis of no spatial dependence in dengue transmission across all space and time combinations, though this evidence reduced as distance between pairs increased and tau approached 1, showing that spatial dependence exists up until a certain distance that was not found in this analysis.

However, one drawback of this analysis is that it required the exclusion of the 67 virologically confirmed dengue cases that had an unknown serotype. This made up 17.4% of the total VCD cases which is a substantial amount. Finding an alternative method may have been useful to incorporate these excluded participants which could have potentially provided some further information in the quantification of spatial dependence. Another drawback of this method is that we have no measure of uncertainty for the actual tau estimates themselves such as a 95% confidence interval as we only obtained point estimates for tau and a 95% confidence interval from the permutation-based null distribution. This means that although we see a clear relationship between proximity of a participant to a VCD case and the tau estimates across all illness onset times, we cannot be sure that this relationship was only observed due to chance as we do not have corresponding 95% confidence intervals for these estimates. Further research could involve using some permutation or bootstrap based method to obtain a measure of uncertainty in order to make statistical inferences about estimates of tau.

5. Cluster Reallocation and Spillover Effects

5.1. Introduction

5.1.1. Spillover Effects

Spillover effects are the effect of an intervention on individuals who are in physical or social proximity to other intervention recipients (22). The difference between spillover effects and the more commonly known term of contamination is that contamination involves individuals unintentionally receiving or being exposed to the intervention whereas spillover affects the individuals who do not receive the intervention; that is, it can occur without individuals in the control arm receiving the intervention (22). Note that the converse also holds true for spillover effects as it can also refer to not receiving the intervention (22). For example, consider a situation in which an individual receives an intervention that directly benefits them and results in spillover. If no other intervention participants are nearby, then nobody will be subject to spillover. However, if other intervention participants are nearby, then they may be affected by spillover from the nearby intervention participants meaning they receive an additional total benefit from their own intervention and the spillover of the interventions from participants nearby (22).

Spillover effects can also be split into positive or negative effects. Here, a positive spillover effect benefits individuals who are affected by the spillover, and a negative spillover effect harms individuals who are affected by the spillover (22). In cluster randomised trials based on location, there is often an assumption of the absence of between-cluster spillover being that people and diseases move freely within a cluster but do not move between clusters (22). Spillover between clusters can occur when clusters are connected through some mechanism (22). For example, in the AWED trial, it is evident that there is potential for spillover effects to be present since each of the clusters are bordering each other as opposed to being in totally separate geographical locations such as schools in a city. Likely, the assumption previously stated that there is no between-cluster movement is almost certainly not true in this context. Valid inferences in randomised experiments rely on the Stable Unit Treatment Value Assumption which requires that differences between the outcomes of control and intervention participants only depend on the participant's intervention status and not on what intervention others receive (22). If spillover effects are present, then this would violate this assumption, even at the cluster level, resulting in biasing the causal effects estimated from randomised comparisons (22). Hence, the identification of possible spillover effects in cluster randomised trials is valuable in order to attempt to take them into account when making causal inferences; even though it adds complexity, the knowledge of the existence and direction of spillover effects may also help in the implementation of intervention strategies in the future.

There is currently no standard method of identifying the existence of spatial spillover, but some approaches have been utilised in the analysis of cluster randomised trials (22). One method includes looking at the distance from the nearest intervention household (22). Another uses the number of surrounding intervention households in the context of insecticide treated net trials (22). However, these methods make additional assumptions in addition to the mechanism of the spillover assumed to be based on distance. The former method assumes only distance to a single nearest intervention household is important and doesn’t account for the number of intervention households nearby and the latter assumes that spillover is based on the density of intervention observations nearby (22).

5.1.2. Cluster Reallocation

Jarvis proposes a method termed “cluster reallocation” that explores the presence of spatial spillover in cluster randomised trials (22). Other than assuming that spillover is based on proximity to cluster boundaries, this method doesn’t make any further assumptions about the spillover mechanism (22).

Cluster reallocation is an iterative method that works by hypothetically reassigning participants to intervention or control arms of the trial based on their proximity to cluster boundaries. The general idea is that the intervention cluster boundaries are “buffered” by a pre-determined distance e.g. 50m such that the participants in the control arm that are within this 50m buffer of the intervention boundary are now reassigned (or reallocated) into the intervention arm in which the estimand of interest is recalculated using the newly defined trial arms. This process is then repeated by for incrementally increasing buffers until some maximum buffer distance. This process is also repeated in the same manner for the control cluster boundaries meaning the control clusters are buffered by some given distance and the participants in the intervention arm that are within this control boundary buffer are reallocated as control participants and the estimand is recalculated, again repeating by increasing the control buffers incrementally. This method provides estimates of the intervention effect for hypothetical spatial definitions of the intervention and control arms, known as ‘buffered estimates’ (22).

When spillover is absent, increases in the intervention or cluster boundaries result in weaker intervention effect and it is hypothesised in (22) that in the absence of spillover, the observed estimate will be either the maximum or minimum estimate compared to the buffered estimates because reallocation of observations to different trial arms will increase similarities between the two arms therefore dilute he effect.

5.2. Methods

Since the use of cluster reallocation requires the use of an estimand of the user's choice, we will use the estimator of tau as defined in Section 4. Specifically, we define the estimand of choice as $\tau(0,100m)$ with difference in illness onset between pairs of participants being within 7 days. This was chosen as in all of the results throughout the project so far in Sections 3 and 4, the strongest effect estimate was observed when pairs of participants had residencies of less than 100m and illness onsets of less than 7 days so we may see the greatest change in estimated tau for different buffer estimates in this space-time window.

Since tau is the estimator of interest, as in Section 4, this required us to first filter out the VCD cases which had a serotype that was unknown. Furthermore, the process of cluster reallocation requires us to utilise the 'sf' package in R in order to calculate the buffers for each cluster. This meant that it was necessary to transform the AWED data set into a spatial data frame in which we also had to change the projection of the coordinate reference system (CRS) to 32749, which is the code that corresponds to Indonesia since this is where the study area is.

Shape files were then loaded into R as a simple features dataset, changed the projection of the CRS to correspond with Indonesia and then merged with the transformed AWED data as this is the basis in which the buffers for each treatment arm can be achieved. Buffers were first added to intervention arms using the st_buffer() command by first filtering the merged dataset such that it only included the intervention arms and then adding a 50m buffer around these clusters. A comparison of the before and after adding this 50m buffer is shown in Figure 9 below.

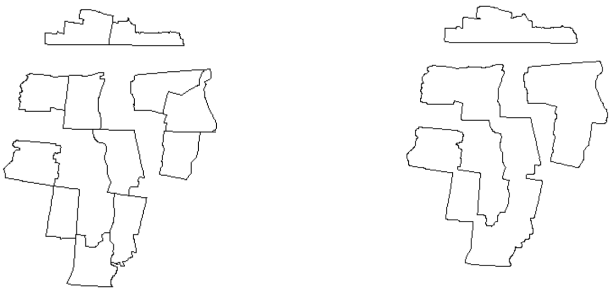

Figure 9: Before and after applying a buffer of 50m to intervention arms

Note that since the control clusters were filtered out, they do not appear in Figure 9 above. A binary vector was then defined by it returning a value of 1 if participants were within these buffered intervention clusters and 0 otherwise. This binary vector was then added to the merged data set as the “new” treatment variable since it defines the reallocated intervention and untreated observations. Then, $\hat{\tau}(0,100m)$ as defined at the start of Section 5.2 was calculated for each arm using these new definitions of intervention and untreated observations. Similar to Section 4, a 95% confidence interval was also obtained through the use of a permutation-based null distribution with 1000 iterations each time using the reallocated observations in order to identify if spatial dependence also existed if clusters were reallocated.

This process was repeated by increasing the buffer of the intervention by 50m each time up until 200m where 200m was chosen as the maximum buffer distance because since each cluster is approximately $1km^2$ in area, by buffering the cluster boundaries more than this amount may have resulted in small numbers of participants in some clusters. Furthermore, this process was also repeated in an analogous way in the untreated arm by buffering initially at 50m and recalculating $\hat{\tau}(0,100m)$ as defined above for both arms and continuing this in 50m intervals up until 200m, also calculating 95% confidence intervals at each step.

5.3. Results

The results of estimating $\hat{\tau}(0,100m)$ for pairs of participants that have illness onset within 7 days for the different buffers for the untreated arm are shown in Figure 10 below.

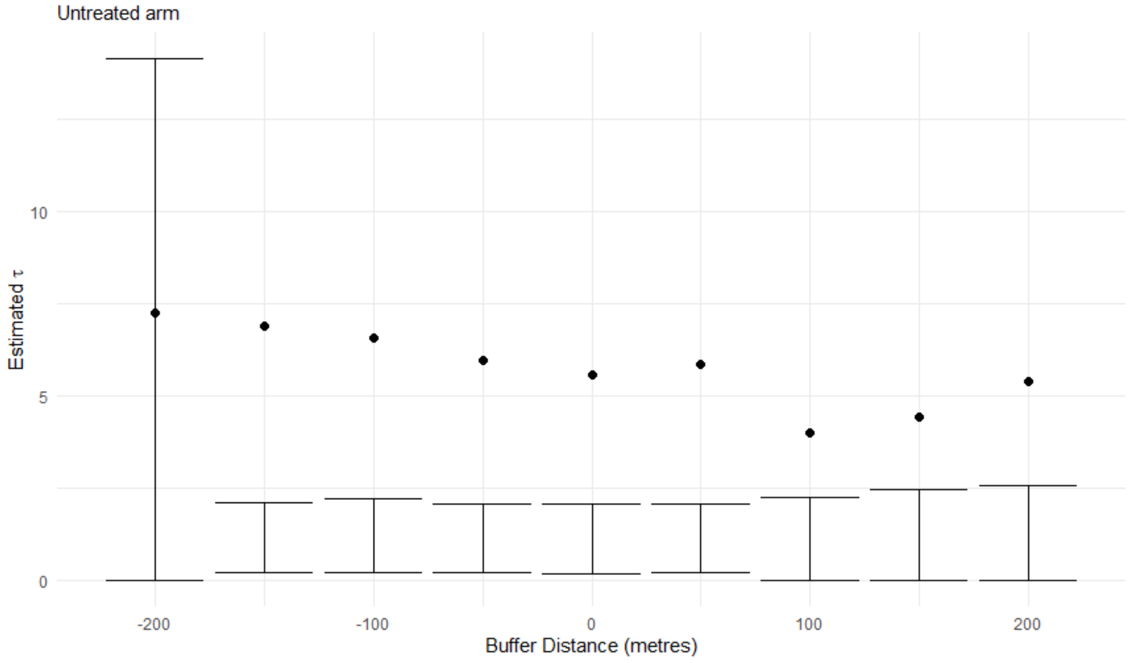

Figure 10: Cluster reallocation plot for $\hat{\tau}(0,100m)$ within 7 days of illness onset in untreated arm

In Figure 10, note that the buffer distance of 0 metres corresponds to the original tau estimate in the untreated arm and a negative buffer distance corresponds to expanding the untreated clusters as opposed to expanding the intervention clusters. For example, a buffer of +100m means to add a buffer of 100m to intervention clusters and then calculate the buffer estimate of tau for the untreated arm; whereas a buffer of -50m means we add a buffer of 50m to untreated clusters and then calculate the buffer estimate of tau for the untreated arm. Here, we see that there is some spillover effect present in the untreated clusters since the original estimate for tau is not the maximum or minimum value out of the buffer estimates.

From Figure 10, the buffer estimates obtained by expanding the intervention clusters are either roughly equal or lower than the original estimate for tau, though this relationship between expanding buffers in the intervention arms and the tau estimates are not monotonic. In this direction, we have that in each case, there is still evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue since each buffer estimate lies outside of the 95% confidence interval given by the permutation-based null distribution obtained from the reallocated participants. However, when the untreated clusters are expanded, we see a constant increase in buffer estimates of tau in comparison to the original estimate. This would make sense as having more participants in the untreated clusters would increase the odds of a participant being a homotypic pair with a case given they live within 100m and had illness onset of less than 7 days of that case relative to the odds of any participant having illness onset of less than 7 days across all untreated areas. We also see that there is evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue up to a buffer of 150m since the buffer estimates up to this point lie outside of the 95% confidence interval. However, when the untreated clusters have a buffer of 200m, we see that there is no evidence to reject this null hypothesis since the buffer estimate at 200m lies within the 95% confidence interval of the null distribution. Note that the 95% confidence interval obtained using this buffer is very wide in comparison to all the other 95% confidence intervals which may be due to an imbalance in the number of participants assigned to each treatment arm following the reallocation process.

The results from estimating $\hat{\tau}(0,100m)$ for pairs of participants that have illness onset within 7 days for the different buffers for the intervention arm are shown in Figure 11 on the next page. Looking at Figure 11, we see that there is little spillover effect in the intervention arm. This is because the original estimate of tau is slightly larger than the buffer estimate of tau that is obtained by buffering the untreated clusters by 50m, showing that it is approximately the minimum value of the buffer estimates. Furthermore, when the intervention clusters are buffered, there is a generally increasing relationship between the buffer and the buffer estimate where the buffer estimate only decreases when the intervention boundaries are buffered by 200m. We also have that there is no evidence of spatial dependence only up until an intervention buffer of 50m and beyond this buffer, we start to see strong evidence to reject the null hypothesis of no spatial dependence in the transmission of dengue since these buffer estimates are far larger than the 95% confidence intervals from the null distribution. This may be due to the intervention clusters now including too many participants from the untreated clusters and they exhibit spatial dependence which would inflate the estimates for tau in the intervention arm. We also have that when expanding the untreated cluster boundaries, there is again a consistent increase in the buffer estimates for tau. Furthermore, in this direction, there is no evidence to reject the null hypothesis of no spatial dependence in dengue transmission for any buffer distance of the untreated clusters since the all the buffer estimates of tau in the intervention arm lie within the 95% confidence interval of the null distribution.

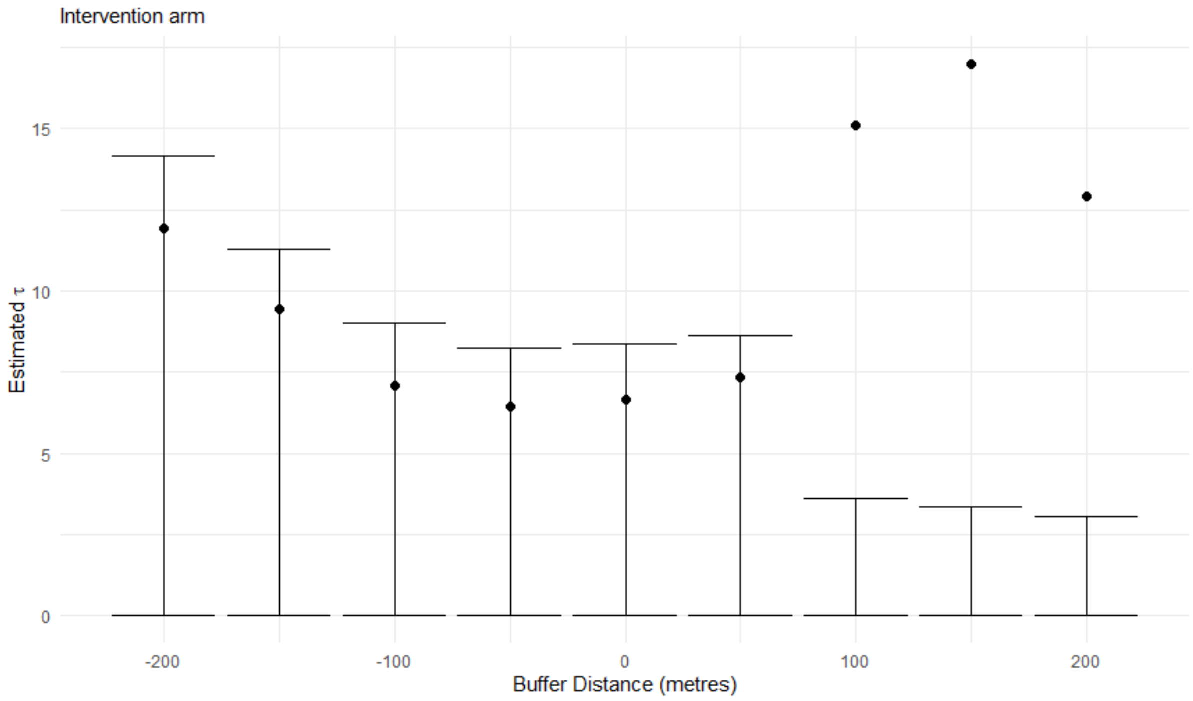

Figure 11: Cluster reallocation plot for $\hat{\tau}(0{,}100m)$ within 7 days of illness onset for intervention arm

5.4. Discussion

In this section, we explored the concept of spatial spillover effects and investigated its existence using the novel method of cluster reallocation. We saw that there was evidence of spillover effects in the untreated arm since the original estimate of $\hat{\tau}(0{,}100m)$ within 7 days of illness onset was neither the maximum nor minimum of the buffer estimates. The cluster reallocation plot in Figure 10 shows a near-linear relationship in the buffer distances and estimated tau which highlights this. This suggests that for those living in the untreated clusters, if they live close to an intervention border, they may experience some of the benefit from being in that cluster as the buffer estimates of tau reduce as the buffer distance for the intervention clusters increase. There was also evidence of spatial dependence for most of the buffer estimates in this treatment arm except when the untreated arm is buffered by 200m. Furthermore, we saw that there was little to no evidence of spillover in the intervention arm since the original tau estimate here was very close to being the minimum value out of the buffer estimates. This suggests that for those living in intervention clusters, there is little to no detriment to living near an untreated border since the buffer estimates for tau relating to the intervention clusters increase regardless of if the intervention or untreated clusters are buffered. We also saw that the spatial independence of the intervention arm is disrupted once the intervention cluster boundaries are buffered from at least 100m.

The method of cluster reallocation has proven useful in demonstrating the presence of spillover and also the buffering distance at which spatial dependence can be seen. One way in which this could be improved is to see how spillover effects and spatial dependence changes for different combinations of tau because one downside of this method is that it is quite computationally intensive meaning it would be time consuming to identify and compare the existence of spatial spillover across different definitions of tau. Since we can currently only identify the existence of spillover through visualisation methods, further research could be untaken in this area to develop a method to quantify the amount of spatial spillover that exists, and therefore conduct statistical inferences on these estimates of spillover through the use of hypothesis tests or confidence intervals.

6. Discussion

6.1. Conclusions

As seen in Section 1, dengue virus is a large public health concern across the world with no universally accepted treatment or prevention scheme. However, the Applying *Wolbachia* to Eliminate Dengue (AWED) trial demonstrated that by infecting the *wMel* strain of the bacteria, *Wolbachia*, into the primary vector of the dengue virus, the *Aedes Aegypti* mosquito, there was a protective efficacy of 77.1% (95% CI: 65.3, 84.9) (16). This cluster randomised trial with test-negative design showed a promising result in which this intervention of *wMel*-infected *Aedes Aegypti* has the scope to be implemented in other study areas in Indonesia or globally to be used as a method to disrupt the transmission of the dengue virus.

One aim of this project was to use the data from the AWED trial to assess cluster-specific and population-averaged effects of exposure and being in an intervention cluster on dengue transmission, and the extent to which the risk of dengue is influenced by proximity of recent dengue cases. These aims were achieved by first giving a definition on what it means for an individual to be exposed. Here, exposure was defined as an individual who had primary residence of at least 10 days between $[d_1, d_2)$ metres of a virologically confirmed dengue case and also had an illness onset with t days of that virologically confirmed dengue case. Contingency tables were first used as a simple analysis between binary exposure and outcome (virologically confirmed dengue) across various times and distances that made up the exposure definition. These results showed that there was a decreasing relationship between the distance between a participant and a VCD case and the estimated odds ratios when the difference in illness onset was within 7 or 14 days. There was an inconsistency when the difference in illness onset was within 30 days as the estimated odds ratios actually increased as the distance between participant and VCD case went beyond 500m. Generalised linear mixed models and generalised estimating equations were used to go beyond the contingency tables by taking clusters into account, and also including intervention status and an interaction between exposure and interaction in both models. The results from these two methods were consistent in that there was strong evidence of an association between exposure and the risk of dengue, controlling for intervention and interaction terms since all space-time combinations resulted in p-values relating to this estimate being p<0.01. This harmful effect of exposure decreases as distance of residence to a VCD case increases, controlling for intervention and interaction terms. Similarly, both methods showed strong evidence of a protective effect of being in an intervention cluster on the risk of dengue, controlling for exposure and interaction terms. This protective effect decreases as proximity to a VCD case increases, with the GLMM only showing weak evidence (p=0.06) of an association between intervention and risk of dengue, controlling for exposure and interaction, when exposure is defined as residing within 1000m of a VCD case and having illness onset within 30 days of that case. There was only evidence of an interaction between exposure and intervention when proximity to a VCD case was set to within 1000m. The main difference between these two methods is that the interpretation of the results from the GLMM only applies to participants of the same cluster which is not very generalisable. However, the results of the GEEs are population-averaged which would be more meaningful in this context.

Another aim of this project was to implement spatio-temporal methods to compare the existence of spatial clustering of dengue between intervention and untreated clusters and how it evolves over time and distance. Section 4 highlighted the use of the estimand $\tau(d_1, d_2)$ which is similar, but not identical to, an odds ratio which measures the ratio of odds between an individual within a treatment arm that enrolled within t days and resides within (d_1, d_2)m of a VCD case in the same arm is a homotypic pair of cases (and therefore potentially transmission related) to the odds that any individuals with illness onset occurring within t days across the whole of that treatment group's areas. Amongst the untreated arm, there was evidence of spatial clustering in the transmission of dengue where differences in time of illness onset between a pair of individuals were within 7, 14, and 30 days and each pair residing up to 1000m from each other. This was evident as each respective estimate of tau calculated for this arm was outside of the 95% confidence interval constructed from a permutation-based null distribution. Amongst the intervention arm, however, there was no evidence to reject the null hypothesis of no spatial clustering in the transmission of dengue for any space-time combination as all of the estimated tau values were within the 95% confidence interval based on the null distribution. Similar to the results from the GLMMs and GEEs, there was a general negative relationship within each arm between the distance that pairs of individuals reside and estimated tau values which was observed for all of the time combinations.

The concept of spillover effect was introduced in Section 5 in which we used the novel method of cluster reallocation to identify if spillover effect exists within each arm. There was evidence of spatial spillover in the untreated arm using the cluster reallocation plot as the original estimate for tau was neither the maximum nor minimum point out of the buffer estimates. A negative trend was observed in which estimates of tau would decrease as we increased the buffer size of the intervention arm. There was very little evidence of spillover in the intervention arm as the original estimate of tau is slightly larger than one of the buffer estimates. It was also evident that this spatial independence within the intervention arm would be disrupted if the intervention arm had a buffer of at least 100m since these buffer estimates were outside of the 95% confidence intervals constructed from the null distribution.

6.2. Limitations and Recommendations

One limitation of the initial trial design that in turn affected the analysis of the project is how the participants are recruited into the trial. Since participants are only recruited if they visit a primary care clinic with symptoms suggestive of dengue fever, this leads to some potential biases. Firstly, as discussed in Section 1.2, approximately 40%-80% of all dengue infections are asymptomatic (7). This is a substantial proportion of potential participants that may have been missed from being included in the trial because they did not have any symptoms at the time. This could mean that there may have been a stronger spatio-temporal dependence on dengue infection than was observed, but as these people did not show symptoms strong enough to warrant a clinic visit, they were not recruited into the trial so the recorded number of VCD cases might have been smaller than the true number. It would be difficult to address this problem from occurring as one method to prevent asymptomatic cases from being missed would be to introduce a regular screening process, but it would be very expensive if every person had to get screened for dengue using the virological or serological tests which can also be seen as invasive as it requires taking a blood sample. Another potential bias of this study design is that there may be selection bias in the enrolment of participants. This is due to the placements of the primary care clinics throughout the Yogyakarta region. Figure 3 shows that there is a slightly uneven spread of primary care clinics which may be the reason why some clusters had more participants than others. For example, cluster 13 has two clinics within the cluster and two other clinics near the border and it also has the largest number of participants. Whereas cluster 24 does not have any primary care clinics nearby and it has one of the lowest numbers of participants which may be due to this lack of accessibility to clinics meaning potential participants might be less likely to be recruited if the clinic is far away. This selection bias could be avoided in the future by placing at least one primary care clinic within each cluster so people can easily access them.

Furthermore, another limitation in the project is that the generalised linear mixed models and generalised estimating equations used in Section 3, unlike the tau estimates used in Sections 4 and 5, do not account for the specific serotype of a VCD case. Here, when looking at risk of being a dengue case, the exposure defined in Section 3.2.1 does not account for serotype information. For example, a participant that is exposed to a VCD case of serotype DENV3 could end up also being a case, but instead have serotype DENV1. Since the serotypes are heterotypic, this pair cannot be transmission related but these models would not account for this. Future analyses could involve using a multinomial logistic regression model with a random effect with 5 outcomes being each of the four dengue serotypes or being a test-negative control and introducing dummy variables for exposure that are 1 if a participant is exposed to a VCD case of a particular serotype, and 0 otherwise.

One of the main limitations of using tau as our estimand for assessing spatial dependence was that it excluded the VCD cases whose serotype was unknown. This is because the estimator is calculated based on pairing individuals based on serotype status so if they are unknown, this is not possible. A sizeable proportion of 17.4% of VCD cases being excluded from this analysis means that there is the potential of loss of information. Since the identification of spillover effects in the treatment arms through the method of cluster reallocation also used tau as the estimand, this means there may have been some loss of information here as well. Another limitation of using tau is that it relies on personal judgement in what space-time combinations are selected which means there is the possibility that potential findings may be lost, such as finding the space-time combination at which spatial dependence is no longer observed in the untreated arm.

Furthermore, recent studies such as (26) describe the use of bootstrap estimates of tau to obtain the null distribution for the case where there is spatial independence and using bias-corrected and accelerated (BCa) 95% confidence intervals instead of percentile-based confidence intervals. This is because BCa confidence intervals can better handle asymmetrical distributions better than percentile-based confidence intervals, and that by using bootstrap tau estimates sampling error was reduced by 24% (26). Although complications arise due to the need to retain spatial structure when resampling using bootstrapping, further investigations into the application of the tau statistic in this trial using this method may yield other insightful results.

References

1. NHS - Conditions - Dengue 2019 [updated 08/08/2019; cited 2021 16/07/2021]. Available from: https://www.nhs.uk/conditions/dengue/.

2. Roy SK, Bhattacharjee S. Dengue virus: Epidemiology, biology, and disease aetiology Canadian Journal of Microbiology. 2021.

3. World Health Organisation - Dengue and Severe Dengue 2021 [updated 19/05/2021; cited 2021 19/07/2021]. Available from: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.

4. World Mosquito Program - Dengue [cited 2021 19/07/2021]. Available from: https://www.worldmosquitoprogram.org/en/learn/mosquito-borne-diseases/dengue.

5. ECDC - Geographical Distribution of Dengue Cases Reported Worldwide, 2020 2021 [updated 21/01/2021; cited 2021 19/07/2021]. Available from: https://www.ecdc.europa.eu/en/publications-data/geographical-distribution-dengue-cases-reported-worldwide-2020.

6. Centers for Disease Control and Prevention - Dengue - Clinical Presentation 2019 [updated 03/05/201915/09/2021]. Available from: https://www.cdc.gov/dengue/healthcare-providers/clinical-presentation.html.

7. European Centre for Disease Prevention and Control - Factsheet about dengue 2021 [updated 04/03/2021; cited 2021 17/09/2021]. Available from: https://www.ecdc.europa.eu/en/dengue-fever/facts.

8. World Health Organisation - Vector-borne Diseases [updated 02/03/2020; cited 2021 20/09/2021]. Available from: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.

9. Powell JR, Tabachnick WJ. History of domestication and spread of *Aedes Aegypti* - A review. Memorias Do Instituto Oswaldo Cruz. 2013.

10. *Aedes Aegypti* - Factsheet for experts [updated 20/12/201620/09/2021]. Available from: https://www.ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/aedes-aegypti.

11. *Aedes Albopictus* - Factsheet for experts [updated 20/12/201621/09/2021]. Available from: https://www.ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/aedes-albopictus.

12. Rather IA, Parray HA, Lone JB, Paek WK, Lim J, Bajpai VK, et al. Prevention and Control Strategies to Counter Dengue Virus Infection Frontiers in Cellular and Infection Microbiology.

13. Aktar MW, Sengupta D, Chowdhury A. Impact of pesticides use in agriculture: their benefits and hazards. Interdisciplinary Toxicology. 2009

14. Lamaningao P, Kanda S, Shimono T, Inthavongsack S, Xaypangna T, Nishiyama T. *Aedes* mosquito surveillance and the use of a larvicide for vector control in a rural area of the Lao People's Democratic Republic. Tropical medicine and health. 2020.

15. *Wolbachia* Method - How it works [21/09/2021]. Available from: https://www.worldmosquitoprogram.org/en/work/wolbachia-method/how-it-works.

16. Utarini A, Indriani C, Ahmad RA, Tantowijoyo W, Arguni E, Supriyati E, et al. Efficacy of *Wolbachia*-Infected Mosquito Deployments for the Control of Dengue. New England Journal of Medicine. 2021.

17. Karyanti MR, Uiterwaal CSPM, Kusriastuti R, Hadinegoro SR, Rovers MM, Heesterbeek H, et al. The changing incidence of Dengue Haemorrhagic Fever in Indonesia: a 45-year registry-based analysis. BMC Infectious Diseases. 2014.

18. Utama IMS, Lukman N, Sukmawati DD, Alisjahbana B, Alam A, Murniati D, et al. Dengue viral infection in Indonesia: Epidemiology, diagnostic challenges, and mutations from an observational cohort study. PLOS Neglected Tropical Diseases. 2019.

19. De Serres G, Skowronski DM, Ambrose CS, Wu XW. The test-negative design: validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. Eurosurveillance. 2013.

20. Szumilas M. Explaining Odds Ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry. 2010;19(3):227-9.

21. Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New York, NY, UNITED STATES: Springer New York; 2006.

22. Jarvis C. Spatial Analysis of Cluster Randomised Trials: London School of Hygiene & Tropical Medicine; Thesis; 2018.

23. Lessler J, Salje H, Grabowski MK, Cummings DAT. Measuring Spatial Dependence for Infectious Disease Epidemiology. PLoS ONE. 2016.

24. Diggle PJ. Spatio-temporal point processes, partial likelihood, foot and mouth disease. Statistical Methods in Medical Research. 2006;15(4):325-36.

25. Dufault SM, Tanamas SK, Indriani C, Utarini A, Ahmad RA, Jewell NP, et al. Disruption of spatiotemporal dependence in dengue transmission by *wMel Wolbachia* in Yogyakarta, Indonesia. MedRxiv. 2021. https://doi.org/10.1101/2021.08.24.21261920

26. Pollington TM, Tildesley MJ, Hollingsworth TD, Chapman LAC. Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic. Spatial Statistics. 2020; 42; 4-6.

Appendix

1. Breakdown of Tau components for different space-time combinations

a	b	c	d	$\hat{\tau}(d_1, d_2)$	d_1	d_2	Treatment	Days
2	1530	40	203290	6.64	0	100	Intervention	7
2	4198	40	203290	2.42	0	200	Intervention	7
2	16028	40	203290	0.63	0	500	Intervention	7
2	25510	40	203290	0.40	0	750	Intervention	7
2	33222	40	203290	0.31	0	1000	Intervention	7
2	2756	58	382714	4.79	0	100	Intervention	14
2	7810	58	382714	1.69	0	200	Intervention	14
4	30122	58	382714	0.88	0	500	Intervention	14
4	48308	58	382714	0.55	0	750	Intervention	14
4	62794	58	382714	0.42	0	1000	Intervention	14
2	5322	146	770342	1.98	0	100	Intervention	30
4	15704	146	770342	1.34	0	200	Intervention	30
10	60410	146	770342	0.87	0	500	Intervention	30
18	96146	146	770342	0.99	0	750	Intervention	30
18	125346	146	770342	0.76	0	1000	Intervention	30
68	1904	1744	272574	5.58	0	100	Untreated	7
106	5070	1744	272574	3.27	0	200	Untreated	7
208	18230	1744	272574	1.78	0	500	Untreated	7
262	29334	1744	272574	1.40	0	750	Untreated	7
330	40046	1744	272574	1.29	0	1000	Untreated	7
112	3462	3154	518808	5.32	0	100	Untreated	14
188	9484	3154	518808	3.26	0	200	Untreated	14
364	34304	3154	518808	1.75	0	500	Untreated	14
464	55398	3154	518808	1.38	0	750	Untreated	14
586	75516	3154	518808	1.28	0	1000	Untreated	14
186	6844	5972	1045336	4.76	0	100	Untreated	30
316	18854	5972	1045336	2.93	0	200	Untreated	30
672	68424	5972	1045336	1.72	0	500	Untreated	30
842	110368	5972	1045336	1.34	0	750	Untreated	30
1084	150828	5972	1045336	1.26	0	1000	Untreated	30