Age and geographic dependence of Zika virus infection during the outbreak on Yap island, 2007

A. Endo¹² and H. Nishiura¹*¹

¹ Graduate School of Medicine, Hokkaido University, Kita 15 Jo Nishi 7 Chome, Kita-ku, Sapporo 060-8638, Japan
² Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, Bloomsbury, London WC1E 7HT, United Kingdom

* Correspondence: Email: nishiurah@med.hokudai.ac.jp; Tel: +81117065066; Fax: +81117067819.

Abstract: Intensive surveillance of Zika virus infection conducted on Yap Island has provided crucial information on the epidemiological characteristics of the virus, but the rate of infection and medical attendance stratified by age and geographical location of the epidemic have yet to be fully clarified. In the present study, we reanalyzed surveillance data reported in a previous study. Likelihood-based Bayesian inference was used to gauge the age and geographically dependent force of infection and age-dependent reporting rate, with unobservable variables imputed by the data augmentation method. The inferred age-dependent component of the force of infection was suggested to be up to 3–4 times higher among older adults than among children. The age-dependent reporting rate ranged from 0.7% (5–9 years old) to 3.3% (50–54 years old). The proportion of serologically confirmed cases among total probable or confirmed cases was estimated to be 44.9%. The cumulative incidence of infection varied by municipality: Median values were over 80% in multiple locations (Gagil, Tomil, and Weloy), but relatively low values (below 50%) were derived in other locations. However, the possibility of a comparably high incidence of infection was not excluded even in municipalities with the lowest estimates. The results suggested a high degree of heterogeneity in the Yap epidemic. The force of infection and reporting rate were higher among older age groups, and this discrepancy implied that the demographic patterns were remarkably different between all infected and medically attended individuals. A higher reporting rate may have reflected more severe clinical presentation among adults. The symptomatic ratio in dengue cases is known to correlate with age, and our findings presumably indicate a similar tendency in Zika virus disease.
Keywords: Zika virus; infectious disease; Asymptomatic case; cumulative incidence; Bayesian inference

1. Introduction

Zika virus is a flavivirus that is genetically linked to dengue virus, yellow fever virus, and West Nile virus. It is mainly transmitted by Aedes mosquitoes. The virus was initially isolated in rhesus monkeys in Uganda in 1947 [1]. Since the identification of the virus, human infection was scarcely observed for decades, with only 14 cases documented before 2007, when the first major outbreak took place on Yap Island, Federated States of Micronesia [2]. This outbreak was followed by larger outbreaks in French Polynesia and three other Oceanian territories in 2013 and 2014 [3–5], and then by the unprecedented widespread epidemic in the Americas in 2015–2016. The first case of the 2015–2016 epidemic was reported in Brazil in March 2015 [6–8]. Since then, circulation of the virus has been confirmed across the world, including 48 countries and territories in the Americas [9]. The causal association of the infection with neurological complications including microcephaly in neonates [10–14] and Guillain–Barré syndrome [13,15,16] raised a global alert regarding this disease.

The clinical presentation of Zika virus disease tends to be mild, and a substantial proportion of infections progress asymptotically, thus concealing the full picture of the disease dynamics. One of the most informative reports on the asymptomatic ratio of Zika virus was produced by Duffy et al. (2009) [2], who presented both case findings and results from household serosurveillance conducted in response to the 2007 epidemic on Yap Island. Their study suggested that, despite the small number of clinically suspected cases (185 patients among 7391 residents), a far larger proportion of the residents had been infected with Zika virus (73.4%) without medical attendance. These results provide an important piece of information on the clinical presentation of Zika virus infection. A serology-based epidemiological study of symptomatic infection indicated that the percent of infected individuals who are symptomatic varies by country and estimated these percentages to be 27% on Yap Island, 44% in French Polynesia, and 50% in Puerto Rico [17], echoing the estimated of 20–27% of pregnant women reported to be symptomatic [18]. Moreover, even provided that symptoms develop, the percentage seeking medical attendance has been estimated to be as low as 7.6–11.5% [19]. Published modelling studies have estimated the reporting rate (i.e., the proportion of infections reported) to be 1.4–1.9% on Yap Island [20] and 7.3–17.9% in French Polynesia [21].

Although overall frequencies of symptomatic infection and reporting have been estimated, heterogeneous patterns of infection and reporting have yet to be clarified. For instance, the age dependency of the cumulative risk of infection and clinical manifestation has not been well studied. Although published reports have illustrated the incidence rate stratified either by age or by geographical location [2,18], the analytical framework has yet to address the age and geographic distribution of cases.

To clarify the heterogeneous characteristics of the epidemic on Yap island, the present study revisited the original study by Duffy et al. (2009) [2], aiming to estimate the age and geographic-specific risk of infection and rate of medical attendance by employing a likelihood-based Bayesian estimation method.
2. Materials and methods

2.1. Data source

We extracted seroprevalence data and the number of confirmed and probable cases of Zika virus infection on Yap Island during the 2007 outbreak [2]. Of a total of 185 patients clinically suspected of having Zika virus disease, the 49 cases that had Zika virus RNA detected in serum or tested serologically positive for Zika virus rather than dengue virus were regarded as confirmed. Cases for which serological testing suggested Zika virus infection but the possibility of dengue virus infection was not sufficiently excluded were considered probable cases (n = 59 cases). We converted the incidence rate by location into case counts by multiplying with the population sizes from the official 2000 census report [22]. The 2000 census was used to calculate incidence rates in the report [2] and thus should reproduce the original case counts. Spatial incidence rates for 10 different locations were also obtained.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Population* [22]</th>
<th>Cases** [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–5</td>
<td>834</td>
<td>6</td>
</tr>
<tr>
<td>5–9</td>
<td>864</td>
<td>5</td>
</tr>
<tr>
<td>10–14</td>
<td>831</td>
<td>8</td>
</tr>
<tr>
<td>15–19</td>
<td>904</td>
<td>5</td>
</tr>
<tr>
<td>20–24</td>
<td>689</td>
<td>8</td>
</tr>
<tr>
<td>25–29</td>
<td>486</td>
<td>6</td>
</tr>
<tr>
<td>30–34</td>
<td>488</td>
<td>15</td>
</tr>
<tr>
<td>35–39</td>
<td>493</td>
<td>13</td>
</tr>
<tr>
<td>40–44</td>
<td>508</td>
<td>9</td>
</tr>
<tr>
<td>45–49</td>
<td>395</td>
<td>9</td>
</tr>
<tr>
<td>50–54</td>
<td>264</td>
<td>15</td>
</tr>
<tr>
<td>55–59</td>
<td>145</td>
<td>8</td>
</tr>
<tr>
<td>60–</td>
<td>490</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>7391</td>
<td>108</td>
</tr>
</tbody>
</table>

* The full breakdown of the population by age and municipality can be found in the supplementary file.
** Probable and confirmed cases.

Table 2. Reported Zika cases by municipality on Yap Island, 2007.

<table>
<thead>
<tr>
<th>Municipality</th>
<th>Population* [22]</th>
<th>Incidence** (per 1000) [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rumung</td>
<td>126</td>
<td>(0.0)</td>
</tr>
<tr>
<td>Maap</td>
<td>592</td>
<td>(11.8)</td>
</tr>
<tr>
<td>Gagil</td>
<td>734</td>
<td>(16.3)</td>
</tr>
<tr>
<td>Tomil</td>
<td>1023</td>
<td>(21.5)</td>
</tr>
<tr>
<td>Fanif</td>
<td>547</td>
<td>(8.4)</td>
</tr>
</tbody>
</table>

Continued on next page
### Table

<table>
<thead>
<tr>
<th>Municipality</th>
<th>Population* [22]</th>
<th>Incidence** (per 1000) [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weloy</td>
<td>1197</td>
<td>(16.7)</td>
</tr>
<tr>
<td>Rull</td>
<td>2019</td>
<td>(11.2)</td>
</tr>
<tr>
<td>Gilman</td>
<td>233</td>
<td>(3.6)</td>
</tr>
<tr>
<td>Kanifay</td>
<td>275</td>
<td>(3.6)</td>
</tr>
<tr>
<td>Dalipebinaw</td>
<td>645</td>
<td>(6.2)</td>
</tr>
</tbody>
</table>

* The full breakdown of the population by age and municipality can be found in the supplementary file.

** Probable and confirmed cases.

### 2.2. A model to estimate the reporting rate

Here, we estimate the reporting rates of infected individuals as a function of age. The reporting rate in different age groups may reflect symptomatic fraction, disease severity and likelihood of seeking medical attendance. Let \( N_{a,g} \) be the population of age group \( a \) in geographical location \( g \). Assuming that the cumulative risk of infection over the entire epidemic period is given as a product of age-dependent relative hazard \( \phi_a \) and spatially specific force of infection \( \lambda_g \), the number of infected individuals \( I_{a,g} \) is considered to follow a binomial distribution:

\[
I_{a,g} \sim \text{Binom}(N_{a,g} \cdot (1 - \exp(-\phi_a \lambda_g)))
\]

where \( 1 - \exp(-\phi_a \lambda_g) \) is the probability that an individual in age group \( a \) and location \( g \) experiences infection. The risk of infection is dependent on both age \( (a) \) and location \( (g) \) because the susceptibility and exposure behaviors vary with age and mosquito transmission potential varies with location.

Let us assume that reporting rate \( p_a \), the probability of an infected individual being reported as a symptomatic case (or more precisely, the probability of being counted as either a confirmed or a probable case), is only dependent on age \( (a) \) because the reporting system is identical across the small island. Moreover, let \( q \) be the diagnostic sensitivity (i.e., the probability of a symptomatic case being correctly classified as a confirmed case, which is assumed to be a biological constant, and thus independent of \( a \) and \( g \)). The number of symptomatic cases of age \( a \) and location \( g \), \( J_{a,g} \), and the total number of confirmed cases of age group \( a \), \( C_a \), can be described as binomial sampling processes:

\[
J_{a,g} \sim \text{Binom}(I_{a,g}, p_a)
\]

\[
C_a \sim \text{Binom} \left( \sum_g J_{a,g}, q \right)
\]

The probability distribution for the results from a seroepidemiological survey (i.e., \( N_{\text{positive}} = 414 \) positives out of \( N_{\text{positive}} + N_{\text{negative}} = 557 \) samples (2)), is described by the hypergeometric sampling process:
We approximated Eq (4) by the binomial distribution with success probability \( \frac{\sum a_g I_{a,g}}{\sum a_g N_{a,g}} \).

\[
N_{\text{positive}} \sim \text{Binom}\left( N_{\text{positive}} + N_{\text{negative}} \frac{\sum a_g I_{a,g}}{\sum a_g N_{a,g}} \right).
\] (5)

Equation (5) is sufficient to quantify reporting rate \( p_a \), which we would like to estimate.

2.3. Statistical analysis

We employed the Markov chain Monte Carlo method [23] combined with data augmentation because we quantified the age dependence from only partial (marginal) information on the observed data: In the empirical observation, \( I_{a,g} \) was not explicitly given and only marginal sums (\( \sum a I_{a,g} \) and \( \sum g I_{a,g} \)) were available, and \( I_{a,g} \) was totally unobserved. The unobserved (or partially observed) variables could be imputed using either the conditional probability function or the Metropolis–Hastings approach. An uninformative prior was employed for all unknown parameters, and median values and 95% credible intervals (CrIs) were derived from 500,000 samples, the first 100,000 of which were discarded as burn-ins.

Given the aforementioned model structure, the parameters could not be uniquely identified because of the high correlation between \( \phi_a \) and \( p_a \). We therefore added smoothing effects to the parameters \( \phi_a \) and \( p_a \) by imposing a constraint on the prior probability. As we expect that the parameter values for similar age groups would also be similar, the prior probability for \( \phi_a \) or \( p_a \) was assumed to be given as a product of two normal distributions according to the adjacent values:

\[
\pi(\phi_a) = \text{Norm}(\phi_{a-1}, r \phi_{a-1})\text{Norm}(\phi_{a+1}, r \phi_{a+1}),
\] (6)

\[
\pi(p_a) = \text{Norm}(p_{a-1}, r p_{a-1})\text{Norm}(p_{a+1}, r p_{a+1}).
\] (7)

When \( a \) corresponds to the youngest or oldest age group, one of the normal distributions in each of Eqs (6) and (7) does not exist; a single normal distribution was used in such cases. The relative standard deviation (RSD) \( r \), the stand deviation of the normal distribution relative to the mean (indicating the strength of the smoothing effect), was set at 1 (i.e., an RSD of 100%) in our baseline analysis. As part of the sensitivity analysis, we also varied the RSD from 50% to 200%.

3. Results

The parameters were estimated using the data augmentation method. Figure 1 shows the medians and 95% CrIs of posterior distributions for the age-dependent relative hazard of infection \( \phi_a \) and reporting rate \( p_a \). Although the uncertainty bounds were broad, Figure 1A indicates that the risk of infection was greater among infants and older adults, and Figure 1B reveals that the reporting rate was highest among adults aged 50–54 years. In general, the trend suggested that children may have been less likely to be reported, compared with adults and older adults. The reporting rate ranged
from 0.7% among those aged 5–9 years to 3.3% among those aged 50–54 years, indicating a substantial percentage of subclinical infections. Figure 1C shows the infected fraction by geographic location. Residents of Tomil exhibited the highest risk of having observed or unobserved Zika virus infections. The overall infected fraction varied by region: In some municipalities, the median was above 80% (Gagil, Tomil, and Weloy), and this value was below 50% in certain other municipalities (Rumung, Gilman, and Kanifay). However, the upper bounds of the CrIs show that the prevalence in those regions might have been higher: The upper bound of the CrI includes the infected fraction of 80% even in Rumung, where no cases were reported during the epidemic. The sensitivity of serological diagnosis \( q \) (i.e., the probability of a clinically suspected case being serologically confirmed given infection) was estimated at 44.9% (95% CrI: 35.8–54.3).

The estimated infected fractions in different geographical locations are displayed as a heatmap in Figure 2A, which illustrates very different demographic patterns compared with the patterns of clinically observed cases shown in Figure 2B. As was the case in Figures 1A and 1B, the frequency of infection and the reporting rate were higher among older age groups. Both were up to 3–4 times higher among adults than among children. Large variation in the age-dependent reporting rate resulted in two different epidemiological landscapes for infected and reported cases on Yap Island, suggesting the existence of a substantial number of unreported infections among children.

**Figure 1.** Inferred epidemiological characteristics of the Zika epidemic on Yap Island, Federated States of Micronesia, in 2007. (A) Age-dependent relative frequency of infection; (B) Age-dependent reporting rate; (C) Estimated cumulative incidence of infection in each municipality. Medians and 95% credible intervals of the posterior distributions are denoted by dots and whiskers.
We varied the RSD of the Gaussian distributions used to smooth the parameters $\phi_\alpha$ and $p_\alpha$ as part of the sensitivity analysis. Changing the RSD from 100% (baseline) to 50% or 200% had limited effects on the age-dependent patterns (Figure 3A–D). It is particularly notable that the age dependence of reporting rate was consistently observed with different RSD assumptions (Figure 3B,D).

**Figure 2.** Heat maps representing inferred age and geographical distribution of infection and reported cases. (A) Number of infected population stratified by age and geographical location; (B) Number of reported cases stratified by age and geographical location. Medians of the posterior distributions are denoted by the color spectrum.

**Figure 3.** Sensitivity analysis to different relative standard deviations (RSD) used for parameter smoothing. Parameters were estimated by varying the RSD of the normal distributions used for smoothing $N_{a.g}$, $\phi_{a_i}$, and $p_\alpha$. (A) and (B) Age-dependent relative frequency of infection and reporting rate with an RSD of 50%; (C) and (D) Age-dependent relative frequency of infection and reporting rate with an RSD of 200%.
4. Discussion

A likelihood-based Bayesian inference framework was applied to documented survey data on the 2007 Zika epidemic on Yap Island. The age and geographical distributions of infected and reported cases were reconstructed using the data augmentation method. The results suggested a heterogeneous epidemiological landscape of Zika virus disease in the 2007 epidemic. Higher incidence of infection was observed among older age groups in the Yap epidemic.

The present study disentangled such age dependency into the combination of two factors—age-dependent frequency of infection and reporting rate given infection—simultaneously accounting for the geographical heterogeneity in the risk of infection. High reporting rates among older age groups imply that the clinical presentation of the disease might have been more severe among adults. A previous study has suggested that age was a risk factor for the severity of Zika virus disease [24]. A correlation between age and symptomatic fraction has also been identified in multiple studies on dengue, a disease that is virologically and clinically similar to Zika virus disease [25–27]. The infection risk has also been suggested to be larger among the older population. Similar findings on the age dependency of dengue fever were found by a serological study in Singapore [28]. Exposure to vectors and biological susceptibility are possible factors responsible for higher risk among older age groups. However, our results were inconclusive because of the wide CrIs, and further epidemiological studies are needed. Nevertheless, we believe it is of interest that our findings suggest that the large number of reported adult cases resulted from the combination of high infection risk and reporting rate. When predicting future disease dynamics, the implication that adults may suffer more severe symptom manifestation may be of note. Specifically, if we assume that Zika virus could become endemic or periodically circulate in some regions, as has been observed in the case of dengue [21], elevation of the age of exposure driven by control efforts or demographic changes [27,29,30] could result in an increased disease burden.

The present study assumed fixed force of infection specific to age and location, and did not explicitly consider interaction, i.e., transmission between individuals. Such models are referred to as static models; the best known static model used in infectious disease modeling is probably the catalytic model, which assumes the force of infection in the population is constant over years and that the age-dependent seroconversion is thus modeled by an exponential decay [31–33]. While catalytic models assume endemic states over multiple years, our model considered single-year force of infection as the dataset we used documented the first introduction of Zika virus to the fully susceptible island. Dynamic models, such as the compartmental susceptible-infectious-recovered (SIR) model [34], are preferred to static models especially when, for example, detailed time-series data is available or the effect of interventions is of interest [35]. However, we believe our simple static model was suitable for the present study to capture the basic epidemiological properties of Zika virus disease in Yap Island.

It should be noted that our analysis rested on multiple assumptions. We assumed that the age dependency of the force of infection was identical across different geographical locations and that the reporting rate was determined solely by age. Therefore, our model might have oversimplified the age and geographically dependent transmission dynamics, which may be more complicated. Additionally, we assumed that the seroprevalence data for Yap Island were generated by an unbiased random sampling of the overall population. We used the Yap population census 2000 in the analysis to be consistent with the original study [2]. Although the total population size in Yap Island
remained stable between 2000 and 2010 (only 1% increase from 11,241 and 11,377), the median age increased by 4 years (20.9 to 25.0) [36]. Use of the 2000 census data as a proxy for the 2007 population might thus have introduced upward bias in the age-dependent trend.

Many aspects of the natural history of infection with Zika virus remain to be clarified. The high asymptomatic ratio and mild clinical presentation (especially among children) mean that most of the transmission dynamics are unobservable; thus, hospital-based surveillance alone can hardly be comprehensive. Carefully designed active surveillance combined with model-based statistical analysis is essential to further clarify the characteristics and behavior of this inconspicuous infectious disease.

5. Conclusion

We applied a Bayesian inference framework to the surveillance data on the 2007 Zika epidemic on Yap Island reported by Duffy et al. (2009) [2] to reconstruct the age and geographical distribution of infected and reported cases. Our model suggests that the high incidence rate among older age groups observed in the Yap epidemic was caused by the frequency of infection and the reporting rate, both of which were higher among older age groups, and that infections in children were significantly underreported. The reporting rate ranged from 0.7% for those aged 5–9 years to 3.3% for those aged 50–54 years, indicating that most infections progressed subclinically. The cumulative incidence of infection was estimated to be over 80% in the municipalities of Gagil, Tomil, and Weloy and below 50% in Rumung, Gilman, and Kanifay. A certain proportion of the population may have been affected even in Rumung, where no cases were reported during the epidemic.

Acknowledgments

HN received funding from the Japan Agency for Medical Research and Development (AMED; grant number: JP18fk0108050), the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant numbers: 16KT0130, 16K15356, 17H04701, 17H05808, and 18H04895), the Inamori Foundation, and the Japan Science and Technology Agency (JST) CREST program (grant number: JPMJCR1413). The funders had no role in study design, data collection and analysis, preparation of the manuscript or the decision to publish. We thank Jennifer Barrett PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

Conflict of interest

All authors declare there is no conflict in this paper.

References


©2020 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)