Epidemics of chikungunya, Zika, and COVID-19 reveal bias in case-based mapping

Fausto Andres Bustos Carrillo (ORCID: 0000-0002-7263-5625)[1]; Brenda Lopez Mercado[2]; Jairo Carey Monterrey[2]; Damaris Collado[2]; Saira Saborio[2]; Tatiana Miranda[2]; Carlos Barilla[2]; Sergio Ojeda[2]; Nery Sanchez[2]; Miguel Plazaola[2]; Harold Suazo Laguna[2]; Douglas Elizondo[2]; Sonia Arguello[2]; Anna M. Gajewski[2]; Hannah E. Maier[3]; Krista Latta[3]; Bradley Carlson[3]; Josefina Coloma[1]; Leah Katzelnick[1]; Hugh Sturrock[4,5]; Angel Balmaseda[2,6]; Guillermina Kuan[2,6]; Aubree Gordon (ORCID: 0000-0002-9352-7877)[3#*]; Eva Harris (ORCID: 0000-0002-7238-4037)[1#*]

Author affiliations:

[1] University of California, Berkeley, Berkeley, California, USA

[2] Sustainable Sciences Institute, Managua, Nicaragua

[3] University of Michigan, Ann Arbor, Michigan, USA

[4] University of California, San Francisco, San Francisco, California, USA

[5] Locational, Poole, UK

[6] Ministry of Health, Managua, Nicaragua

# These senior authors contributed equally to this article

***Co-corresponding authors:**

Eva Harris, Division of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, 185 Li Ka Shing Center, 1951 Oxford Street, Berkeley, CA 94720-3370; Tel. 1-510-642-4845; eharris@berkeley.edu

Aubree Gordon, Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, 5622 SPH I, 1415 Washington Heights, Ann Arbor, MI 48109-2029; Tel. 1-510-409-5495; gordonal@umich.edu

Main text: 3500 / 3500, Abstract: 150 / 150

Keywords: Chikungunya, Zika, COVID-19, bias, spatial analysis, epidemiology

Running title: Four epidemics reveal bias in case-based mapping

Article summary line: Inferring measures of spatial risk from case-only data can substantially bias estimates, thereby weakening and potentially misdirecting measures needed to control an epidemic.

ABSTRACT

Accurate tracing of epidemic spread over space enables effective control measures. We examined three metrics of infection and disease in a pediatric cohort (N≈3,000) over two chikungunya and one Zika epidemic, and in a household cohort (N=1,793) over one COVID-19 epidemic in Managua, Nicaragua. We compared spatial incidence rates (cases/total population), infection risks (infections/total population), and disease risks (cases/infected population). We used generalized additive and mixed-effects models, Kulldorf's spatial scan statistic, and intracluster correlation coefficients. Across different analyses and all epidemics, incidence rates considerably underestimated infection and disease risks, producing large and spatially non-uniform biases distinct from biases due to incomplete case ascertainment. Infection and disease risks exhibited distinct spatial patterns, and incidence clusters inconsistently identified areas of either risk. While incidence rates are commonly used to infer infection and disease risk in a population, we find that this can induce substantial biases and adversely impact policies to control epidemics.

INTRODUCTION

Controlling epidemic spread requires accurate data on the movement of pathogens through populations. Standard spatial studies of infectious diseases use passively collected, individual-level data for cases (symptomatic infections) from health facilities after cases present for medical treatment (1–5). Then, by using census data to obtain the total population in an area, these studies estimate incidence rates (attack rates, incidence proportions) as the ratio of cases to the total population. However, because this approach does not capture subclinical (clinically inapparent) infections, this incidence approach may not recapitulate the spatial contour of infections, which may have a distinct pattern and magnitude. These issues may be compounded when the incidence rate, estimated from passively collected and hence incomplete case data, is used to infer infection risk (1–3) or disease risk (4,5) in policy decision-making on epidemic control.

Epidemiological risk is the probability of a susceptible individual experiencing an outcome. For an immunologically naïve population, all persons are at risk for an initial infection. However, only infected individuals are at risk for experiencing illness, as only infected persons are susceptible to disease. Consequently, measuring infection status is necessary to estimate the numerator of infection risk (infections/total population) and the denominator of disease risk (cases/infected population). These metrics are related to the incidence rate through an application of conditional probability, expressed in multiple ways below:

Epidemiological: $$\text{Infection risk} \times \text{Disease risk} = \text{Incidence rate} \quad \text{(Eq. 1)}$$

Algebraic: $$\frac{\text{Infections}}{\text{Total population}} \times \frac{\text{Cases}}{\text{Infections}} = \frac{\text{Cases}}{\text{Total population}}$$

Statistical: $$P(\text{Infection}) \times P(\text{Disease} \mid \text{Infection}) = P(\text{Disease and Infection})$$

Eq. 1, applicable to infectious disease epidemics in an initially naïve population, demonstrates that incidence is the product of two underlying probabilities of interest. Thus, the incidence rate is explained by, and can be decoupled into, infection and disease risks.

We spatially analyzed four explosive epidemics in two longitudinal Nicaraguan cohorts. Our analysis covers the 2014 and 2015 chikungunya epidemics caused by chikungunya virus (CHIKV) (6,7), the 2016 Zika epidemic caused by Zika virus (ZIKV) (8,9), and the first wave of the COVID-19 epidemic in 2020 caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (10). While *Aedes* mosquitoes transmit CHIKV and ZIKV (11), SARS-CoV-2 primarily spreads by respiratory droplets (12). We analyzed the epidemics in parallel to identify commonalities across epidemics of different pathogens and transmission routes. We demonstrated differences in the fine-scale spatial characterization of epidemics by standard incidence-based measures versus a more comprehensive approach that included infection and disease risks. Finally, we quantified and mapped the separate biases induced by using passive versus active surveillance.

METHODS

Ethics statement

The Pediatric Dengue Cohort Study (PDCS) was approved by Institutional Review Boards (IRBs) of the University of California, Berkeley; the University of Michigan, Ann Arbor; and the Nicaraguan Ministry of Health. The Household Influenza Cohort Study (HICS) was approved by the University of Michigan, Ann Arbor, and the Nicaraguan Ministry of Health IRBs. Participants' parents or legal guardians provided written informed consent. Subjects six years and older provided verbal assent.

Study design and eligibility criteria

The PDCS (13) is an open, population-based, prospective cohort of children initiated in 2004 to study dengue virus and later expanded to include CHIKV and ZIKV. We assessed ~3,000 PDCS participants 2-14 years old who experienced two chikungunya epidemics and one Zika epidemic (6–9). The HICS is an open, population-based, prospective cohort of households that has studied influenza virus and coronaviruses since 2017. We evaluated 1,793 HICS participants 0-87 years old who experienced the first COVID-19 epidemic (10). The age structure of the HICS is representative of Managua's general population.

Both cohort studies share the same study site (Fig. 1) in Managua, Nicaragua's capital. During the studies' annual sampling (serosurvey) in March/April, participants provide blood samples to ascertain infection status during the prior year. A mid-year sampling was instituted in the HICS in October/November 2020 to measure SARS-CoV-2 infections after the first COVID-19 wave but before the second. Both studies provide participants with primary care; participants agree to visit the study health center at the first indication of any illness.

Analysis of each epidemic was restricted to participants who lived within the health center's catchment area and were immunologically naïve. By further restricting to participants who were enrolled before each epidemic, we analyzed a closed cohort of initially uninfected participants who subsequently experienced an epidemic. The Appendix (pages 1-3) contains additional study design information.

Laboratory methods

Upon collection, annual blood samples were immediately transported to the Nicaraguan National Virology Laboratory for processing and storage at -80°C. Paired annual samples (2014-2015 and 2015-2016) demonstrating seroconversion by CHIKV Inhibition ELISA (14) indicated

CHIKV infection. ZIKV infection status was confirmed by the 2017 result of the ZIKV NS1 blockade-of-binding assay (15) on paired 2017-2018 annual samples. SARS-CoV-2 infection status was confirmed by the "Mount Sinai ELISA" protocol (16), primarily on 2020 midyear samples. Participants with laboratory-confirmed infections who did not seek medical care were categorized as experiencing subclinical infections. Acute and convalescent samples from participants suspected of chikungunya, Zika, or COVID-19 were tested using molecular, virological, and serological assays (7,8,13). The Appendix (pages 3-4) contains detailed laboratory methods.

Statistical analyses

We measured the incidence rate, infection risk, and disease risk of each epidemic. Overall values of these metrics were estimated using intercept-only logistic models. The metrics' values across the study area were estimated with generalized additive models (17) using two-dimensional splines on households' longitude and latitude, where participants were geolocated. To quantify bias arising from incomplete case ascertainment, Zika case data was disaggregated by whether they were obtainable through active or passive surveillance, the only epidemic where this was possible. The intracluster correlation coefficient was used to measure the intra-household correlation of infection and disease outcomes. We used SaTScan v9.4.4 and Kulldorf's spatial scan statistic to identify hierarchical and Gini clusters of case incidence, infection risk, and disease risk (18,19). Geostatistical mixed models (20) were used to describe the association of risk factors with infection and disease outcomes. Infection dynamics were estimated by treating cases as a spatiotemporal Poisson point process arising from the total population and then accounting for the spatial distribution of disease risk, assumed to be time-invariant. Initially uninfected participants were considered at risk for infection; infected

participants were considered at risk for disease. Analyses used the EPSG:4326 coordinate reference system and were performed in R v3.6.2. The Appendix (pages 5-15) contains detailed statistical methods.

RESULTS

Participant characteristics

We refer to the first chikungunya epidemic as *ChikE1*, the second as *ChikE2*, the Zika epidemic as *ZikaE,* and the COVID-19 epidemic as *CovidE*. Our study assessed infection and disease outcomes for 4,884 distinct individuals, including 3,693 unique PDCS participants across ChikE1, ChikE2, and ZikaE. Of the 1,793 HICS participants, 602 children were also enrolled in the PDCS, and 1,192 mostly adult participants were only enrolled in the HICS. Approximately 3,000 PDCS participants were analyzed in ChikE1, ChikE2, and ZikaE (Table 1). These three epidemics occurred in 2014-2016 throughout Managua's rainy period of June-November (Fig. 2), during which an abundance of mosquitoes is observed in the study area. In contrast, CovidE peaked during May-July of 2020.

In the PDCS, the distribution of age and sex was constant across ChikE1, ChikE2, and ZikaE (Table S1, Fig. S1), with approximately 50% of PDCS participants being female. In the HICS, there was an over-enrollment of adult females relative to adult males.

Summary measures of infection, disease, and case-based incidence

We first examined summary statistics of the four epidemics. ChikE1 exhibited the lowest incidence at 2.9 cases per 100 population (2.9%) but featured higher infection (6.4%) and disease risks (45.6%) (Table 1). ChikE2, ZikaE, and CovidE exhibited similar incidence rates between 14.5-17.1%, but these incidence rates differed substantially from infection and disease risks.

ChikE2 had a medium level of infection risk (24.8%) and a high disease risk (58.7%), an inverted pattern from what was observed during CovidE, with high infection risk (57.5%) and medium disease risk (28.9%) (Table 1). In contrast, ZikeE displayed intermediate levels of infection (47.1%) and disease (35.4%) risk. Across epidemics, the case-based incidence rate thus recapitulated neither risk-based metric and often underestimated them considerably (Table 1).

We then assessed summary statistics by sex and age. Sex-based differences for incidence and risk-based measures, even when statistically significant, tended to be small, as when females had an infection risk 6% higher than males during ZikaE (Fig. S2). Similarly, accounting for the over-enrollment of adult females in the HICS had little effect (~1%) on overall estimates (Fig. S3-S4). In contrast, we observed age-based incidence patterns for all epidemics (Fig S5), which were explained by the underlying and more striking age trends observed for infection and/or disease risks (Fig S6-8). For example, COVID-19 incidence was low across age, particularly during childhood (Fig. S5). However, SARS-CoV-2 infection risk was high across all ages, increasing modestly from ~48% in infants to ~62% at age 24 and thereafter plateauing. Despite the relative stability of infection risk by age, disease risk during CovidE increased dramatically from ~11% in infants to ~50% at age 70. Thus, the low COVID-19 incidence neither recapitulated age-based risk dynamics nor reflected the greater age-based changes in disease risk as compared to infection risk.

Mapping infection, disease, and incidence

Next, we mapped the infection risk, disease risk, and case-only incidence rate across our study area. For all epidemics, infection risk varied at small spatial scales (Fig. 3A-D), suggesting that the local environment was an important determinant of infection risk. During ChikE1, ChikE2, and ZikaE, infection risk was elevated in western neighborhoods adjacent to a large

cemetery that is heavily infested with *Aedes* mosquitoes during the rainy season (data not shown). Only adjusting Fig. 3A-D for distance to the cemetery appreciably changed the spatial patterns of infection risk, whereas adjusting for age, sex, and household water availability did not (Figs. S9-S14). Conversely, SARS-CoV-2 infection risk was high in eastern neighborhoods that contain large public spaces and commercial attractions (Figs. 3D, S15-S16). Together, these observations imply that infection risk across epidemics was spatially mediated by distinct transmission routes.

Across all epidemics, disease risk also varied at small spatial scales (Fig. 3E-H). After adjusting for age and sex (Figs S17-24), spatial patterns of disease risk remained non-uniform and distinct from spatial patterns of infection risk. This demonstrates that disease risk can vary spatially and that areas of high infection risk may not have commensurate levels of disease risk.

As the case-based incidence rate is the product of two risks (Eq. 1) with different spatial patterns (Fig. 3A-H), maps of the incidence rate (Fig. 3I-L) underestimated infection and disease risk-based maps and did not recapitulate spatial patterns of either risk. We quantified the bias resulting from treating the incidence rate as infection and disease risk by subtracting incidence maps from risk-based maps (Fig. 3M-T). The average spatial bias for the disease risk was -40.1 and -41.2 percentage points for ChikE1 and ChikE2, respectively; the average spatial bias for the infection risk was -34.9 and -40.3 percentage points for ZikaE and CovidE, respectively. The incidence rate underestimates risk-based metrics, inducing negative biases. Additionally, bias varied substantially across neighborhoods. For example, the *range* of bias for ChikE2 and CovidE infection risks was 31.2 and 19.2 percentage points across the study area. Thus, the inferential bias induced by treating the incidence rate as a risk-based metric was high and spatially heterogeneous across epidemics.

Cluster detection

We then identified hierarchical and Gini clusters of infection risk, disease risk, and incidence (Fig. 4, Table S2). Each epidemic had ≥1 significant infection or disease cluster. Clusters of elevated infection risk for the larger mosquito-borne epidemics, ChikE2 and ZikaE, encompassed the cemetery and study neighborhoods adjacent to it. Large clusters of diminished infection risk in ChikE1, ChikE2, and ZikaE highlighted areas with excess uninfected persons who remained susceptible to future infection. Such clusters are only identifiable after ascertaining the infection status of a population, regardless of disease presentation. In contrast to the mosquito-borne epidemics, CovidE exhibited small clusters of elevated and diminished infection risk. In general, clusters of infection risk were in different locations and of different sizes than clusters of disease risk, demonstrating that infection and disease risk cluster differently in space. Indeed, we detected no disease risk clusters during both chikungunya epidemics despite finding large clusters of infection risk.

Standard incidence clusters, which identify areas of elevated or diminished case counts among the total population, sometimes missed large risk-based clusters (Fig. 4). More surprisingly, incidence clusters resembled infection risk clusters only for ChikE2 and CovidE, whereas they resembled disease risk clusters for ChikE1 and ZikaE. Thus, incidence clusters failed to display a reproducible pattern, inconsistently resembling either infection or disease risk clusters for a given epidemic.

Geostatistical modeling

We next conducted geostatistical multivariable modeling. We first describe model-based inferences for correlated outcomes within households and across space. Surprisingly, analyses that did and did not account for household-based correlation yielded very similar results for all

234 epidemics, suggesting that participants' infection and disease outcomes were poorly correlated
235 within homes (Tables S3-S4). This observation was directly confirmed by low values of the
236 intracluster correlation coefficient (Tables 1, S5). We further observed that infection risk did not
237 scale with household size (Fig. S25). Altogether, the data demonstrated that our participants'
238 infection and disease outcomes were weakly correlated within households across epidemics.
239 Likewise, the similarity of estimates from models that did and did not account for spatial
240 autocorrelation (Tables S3-S4) suggested that infection and disease outcomes were only spatially
241 correlated across short distances. This observation was confirmed by estimated Matérn
242 correlation functions (Fig. S26) that demonstrated that infection and disease outcomes were
243 spatially correlated across short distances (<200m) for all epidemics, strengthening earlier
244 findings (Fig. 3) regarding the importance of the local spatial environment.
245 Indeed, we observed that distance to the cemetery was significantly associated with
246 ZIKV *infection*, such that the odds of ZIKV infection among participants living 1 km from the
247 cemetery were 0.63 (95% CI: 0.55, 0.73) times that of participants living next to the cemetery,
248 conditional on age, sex, and indoor water availability; a similar 1-km odds ratio was observed
249 during ChikE2 (Table S3). However, using geostatistical models, we did not identify variables
250 that were consistently related to *disease* risk across epidemics (Table S4). Rather, model results
251 were epidemic-specific.

252 **Spatiotemporal dynamics**

253 Spatiotemporal analyses depict epidemic progression across time and space. By
254 harnessing Eq. 1, we estimated the spatiotemporal dynamics of infection risk (Fig. 5A-D), which
255 were substantially underestimated by the less dynamic standard spatiotemporal patterns of case
256 incidence (Fig. 5E-H). Each of the mosquito-borne epidemics featured elevated infection risk

around cemetery-adjacent neighborhoods for ≥2 months, with particularly high infection risk during ZikaE. In contrast, cemetery-adjacent neighborhoods were never the focal point of SARS-CoV-2 infection risk. Because CovidE featured the lowest disease risk (Table 1), its spatiotemporal *infection* dynamics differed most from its *incidence* dynamics, underscoring the substantial differences of risk-based mapping compared to case-based mapping.

Active versus passive surveillance

We quantified case ascertainment bias spatially by complementing the ZikaE serosurvey with either all Zika cases or only cases captured by passive surveillance. Compared to our active surveillance, using passive surveillance altered the clinical profile of captured Zika cases (8) and decreased the case count, thereby increasing the number of subclinical ZIKV infections. The infection risk was unbiased under passive surveillance as its calculation only required serosurvey data; however, estimates of the disease risk and incidence rate were biased (Table 2). The bias from passive surveillance is conceptually and numerically distinct from that induced by treating the incidence rate as a risk. However, these two biases synergized when the incidence rate, estimated from passive surveillance data, was interpreted as a risk. For example, inferring the true disease risk from the incidence rate induced -4.9 percentage points of bias from incorrect inference *and* -26.1 percentage points of bias from incomplete case ascertainment (Table 2). Importantly, this compounded bias would be present irrespective of conducting a serosurvey (Table 2). Moreover, whether biases arose from misinterpretation, incomplete case data, or both, they tended to be high and spatially heterogenous (Figure 6). Thus, inferring risk from passive surveillance data was prone to multiple biases with different spatial patterns.

DISCUSSION

Across multiple analyses and four epidemics of three viruses in two cohorts, we showed that the traditional case-based incidence rate considerably underestimated infection and disease risks, broadly impacting how epidemics were characterized. We further demonstrated that case-based analyses did not recover either the magnitude or spatial pattern of infection risk, which critically conveys the landscape of natural immunity. In general, we observed that case-based incidence had more limitations than traditionally assumed. For example, although ChikE2, ZikaE, and CovidE had comparable incidence rates, their underlying infection and disease risks were very different. Similarly, case-based incidence clusters inconsistently captured different risks across epidemics, an observation not apparent without analyzing multiple epidemics in parallel. Together, our results demonstrate how complex, epidemic-specific spatial patterns of infection and disease risk, critical for the design of effective interventions, can be obscured and underestimated by relying solely on case-based analyses. Importantly, this underestimation was distinct from bias due to incomplete case ascertainment, suggesting that the inferential biases we quantify for the incidence rate are exacerbated in typical settings with limited active surveillance and laboratory testing capacity.

Paradoxically, the limitations of the incidence rate are obvious yet underappreciated. It is well-known that incidence estimates based on incomplete case data are underestimated. Here, we showed that a separate bias, with its own spatial pattern, arises when the incidence rate is misinterpreted as conveying infection or disease risks, and we quantified the extent to which this biased estimate deviates from more accurate estimates of infection and disease risk. Correctly interpreting measures of epidemic impact is important for policy decisions. While interventions will vary depending on the pathogen and available countermeasures, areas prone to high

infection risk generally require interventions that limit transmission (*e.g.*, mosquito control, masking, and social distancing), whereas areas prone to high *disease* risk require interventions that limit disease occurrence and boost access to care.

Incidentally, if the disease risk were spatially uniform, as some studies have assumed (21), then the spatial pattern of incidence would equal that of the infection risk and the degree of underestimation (and hence bias) would be similar across a given area. However, disease risk was not spatially uniform across epidemics, and its bias also varied spatially. Thus, just as others have found that disease risk can vary across populations (6,22), we find that disease risk can vary within a single population.

The case-based incidence rate is the disease risk when all individuals are susceptible to an outcome (*e.g.*, cardiovascular disease, death). However, for pathogens that cause subclinical infections, incidence rate maps only convey *where disease occurred*, not the *spatial risk* of infection or disease. Many pathogens of global health importance give rise to substantial quantities of subclinical infections (including *Plasmodium*; *Mycobacterium tuberculosis*; and many pathogens transmitted by sex, air, vectors, and soil). Thus, our findings concerning the limitations of case-based spatial mapping likely generalize to many infectious diseases that disproportionately affect neglected populations.

The pediatric nature of the PDCS precluded spatially analyzing adults in the catchment area of the study health center during ChikE1, ChikE2, and ZikaE. However, previous analyses compared ZIKV infection risk for children and adults in this area (9). The two groups' spatial patterns were comparable, although ZIKV infection risk was higher among adults. Thus, analyses of the adult population during ChikE1, ChikE2, and ZikaE would likely reveal similar spatial trends as those in PDCS participants.

We found little evidence that infections were clustered within households. Lacking entomological data, our analyses indirectly suggested that viral transmission infrequently occurred within study households. However, this suggestion is directly supported by a study of full-length sequencing of ZIKV genomes in our cohort (23), which found that many households had Zika cases whose most recently sampled viral ancestral strains derived from different households. Together, the evidence suggests that non-household transmission played an important role in the epidemics we assessed.

The geographic extent of our study is small. However, capturing all infections and cases, and hence accurately measuring bias, is only cost-feasible in constrained geographical areas. Spatial studies with incomplete infection and case data, whether small or large, may be subject to inferential and case ascertainment biases despite being unable to measure such biases. Ascertaining infection status and exhancing case surveillance, where possible, may help to mitigate and correct for such biases.

Measuring a population's infection status has many additional benefits, especially in directing infection control interventions to areas of high transmission. Conversely, knowledge of areas with a high proportion of uninfected individuals is also critical for advancing public health goals, such as prioritizing these areas for epidemic-preventive measures (*e.g.*, vaccine rollout in areas with low SARS-CoV-2 transmission). Others have shown how combining regional serosurvey data with real-time hospitalization data can estimate infection risk in near real-time at larger spatial scales, thereby improving critical estimates for decision-makers (24). As epidemic management necessitates evaluating the risks of infection and disease across space, our data supports the expanded use of serosurveys to overcome the inherit limitations of case-based spatial measures.

ACKNOWLEDGMENTS

We are extremely appreciative of our dedicated study team at the Centro de Salud Sócrates Flores Vivas and the Laboratorio Nacional de Virología at the Centro Nacional de Diagnóstico y Referencia, Nicaraguan Ministry of Health; and the Sustainable Sciences Institute in Nicaragua. We are grateful to Art Reingold and Tulika Singh for their thoughtful reviews of the manuscript, and we thank Burke Bundy and Suzanne Default at the University of California, Berkeley, for enabling us to use the computer cluster of the Division of Epidemiology and Biostatistics for our geostatistical modeling. We thank François Rousset for expert consultation regarding spatial generalized linear mixed models and their implementation in the spaMM R package. Most importantly, we thank the PDCS and HICS study participants and their families for engaging with us in the endeavor of science. This study was supported by grants R01 AI099631 (AB), P01 AI106695 (EH), R01 AI120997 (AG), and U19 AI118610 (EH) from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health; the National Institutes of Health Centers of Excellence for Influenza Research and Surveillance [contract: HHS 272201400006C (AG)]; and the Open Philanthropy Project Fund for the production of recombinant SARS-CoV-2 spike protein, its receptor binding domain, and antibodies at the University of Michigan Center for Structural Biology. FBC was partially supported by a supplement to grant P01 AI106695.

ABOUT THE AUTHOR

Dr. Fausto Andres Bustos Carrillo is an epidemiologist in Washington, DC, collaborating with the National Institute of Allergy and Infectious Diseases as an Emerging Leader in Data Science Fellow. His primary research interests are the epidemiological, clinical, and spatial aspects of explosive epidemics caused by SARS-CoV-2, Zika virus, chikungunya virus, dengue virus, and

372 influenza virus, which were the topics of his doctoral dissertation at the University of California,

373 Berkeley.

REFERENCES

1. McHale T, Romero-Vivas C, Fronterre C, Arango-Padilla P, Waterlow N, Nix C, et al. Spatiotemporal heterogeneity in the distribution of chikungunya and Zika virus case incidences during their 2014 to 2016 epidemics in Barranquilla, Colombia. Int J Environ Res Public Health. 2019 May 2;16(10).

2. Aguiar BS, Lorenz C, Virginio F, Suesdek L, Chiaravalloti-Neto F. Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. Int J Infect Dis. 2018 May 1;70:20–9.

3. Campos de Lima EE, Gayawan E, Baptista EA, Queiroz BL. Spatial pattern of COVID-19 deaths and infections in small areas of Brazil. PLoS One. 2021 Feb 1;16(2 Febuary).

4. Vissoci JRN, Rocha TAH, Silva NC da, de Sousa Queiroz RC, Thomaz EBAF, Amaral PVM, et al. Zika virus infection and microcephaly: Evidence regarding geospatial associations. PLoS Negl Trop Dis. 2018 Apr 25;12(4).

5. Bonilla-Aldana DK, Bonilla-Aldana JL, García-Bustos JJ, Lozada CO, Rodríguez-Morales AJ. Geographical trends of chikungunya and Zika in the Colombian Amazonian gateway department, Caqueta, 2015–2018 – Implications for public health and travel medicine. Travel Med Infect Dis 2020 May 1;35.

6. Bustos Carrillo F, Collado D, Sanchez N, Ojeda S, Lopez Mercado B, Burger-Calderon R, et al. Epidemiological evidence for lineage-specific differences in the risk of inapparent chikungunya virus infection. J Virol. 2019 Nov 21;93(4):e01622-18.

7. Gordon A, Gresh L, Ojeda S, Chowell-Puente G, Gonzalez K, Sanchez N, et al. Differences in transmission and disease severity between two successive waves of

chikungunya. Clin Infect Dis. 2018 Apr 25;67(11):1760–7.

8. Burger-Calderon R, Bustos Carrillo F, Gresh L, Ojeda S, Sanchez N, Plazaola M, et al. Age-dependent manifestations and case definitions of paediatric Zika: a prospective cohort study. Lancet Infect Dis. 2020 Dec;20(3):371–80.

9. Zambrana JV, Bustos Carrillo F, Burger-Calderon R, Collado D, Sanchez N, Ojeda S, et al. Seroprevalence, risk factor, and spatial analyses of Zika virus infection after the 2016 epidemic in Managua, Nicaragua. Proc Natl Acad Sci. 2018 Sep 11;115(37):9294–9.

10. Maier HE, Kuan G, Saborio S, Carrillo Bustos F, Plazaola M, Barilla C, et al. Clinical Spectrum of Severe Acute Respiratory Syndrome Coronavirus 2 Infection and Protection From Symptomatic Reinfection. Clin Infect Dis. 2021 Aug 19.

11. Higgs S, Vanlandingham D, Powers A, editors. Chikungunya and Zika Viruses: Global Emerging Health Threats. 1st ed. Academic Press; 2018.

12. Cevik M, Kuppalli K, Kindrachuk J, Peiris M. Virology, transmission, and pathogenesis of SARS-CoV-2. BMJ. 2020 Oct 23;371.

13. Kuan G, Gordon A, Aviles W, Ortega O, Hammond SN, Elizondo D, et al. The Nicaraguan Pediatric Dengue Cohort Study: Study design, methods, use of information technology, and extension to other infectious diseases. Am J Epidemiol. 2009 Jul 1;170(1):120–9.

14. Saborío Galo S, González K, Téllez Y, García N, Pérez L, Gresh L, et al. Development of in-house serological methods for diagnosis and surveillance of chikungunya. Rev Panam Salud Publica. 2017 Aug 21;41:e56.

15. Balmaseda A, Stettler K, Medialdea-Carrera R, Collado D, Jin X, Zambrana JV, et al.

418 Antibody-based assay discriminates Zika virus infection from other flaviviruses. Proc Natl
419 Acad Sci 2017 Aug 1;114(31):8384–9.

420 16. Amanat F, Stadlbauer D, Strohmeier S, Nguyen THO, Chromikova V, McMahon M, et al.
421 A serological assay to detect SARS-CoV-2 seroconversion in humans. Nat Med. 2020 Jul
422 1;26(7):1033–6.

423 17. Hastie T, Tibshirani R. Generalized Additive Models. New York: Chapman & Hall; 1990.

424 18. Kulldorff M. A spatial scan statistic. Commun Stat - Theory Methods. 1997;26(6):1481–
425 96.

426 19. Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb DG, Tatalovich Z, et al. Using Gini
427 coefficient to determining optimal cluster reporting sizes for spatial scan statistics. Int J
428 Health Geogr. 2016 Aug 3;15(1):27.

429 20. Rousset F, Ferdy J-B. Testing environmental and genetic effects in the presence of spatial
430 autocorrelation. Ecography (Cop). 2014;37(8):781–90.

431 21. Moore SM, ten Bosch QA, Siraj AS, Soda KJ, España G, Campo A, et al. Local and
432 regional dynamics of chikungunya virus transmission in Colombia: The role of
433 mismatched spatial heterogeneity. BMC Med 2018 Aug 30;16(1):152.

434 22. Haby M, Pinart M, Elias V, Reveiz L. Prevalence of asymptomatic Zika virus infection: a
435 systematic review. Bull World Health Organ. 2018 Jun 1;96(6):402-413D.

436 23. Sun H, Binder RA, Dickens B, de Sessions PF, Rabaa MA, Ho EXP, et al. Viral genome-
437 based Zika virus transmission dynamics in a paediatric cohort during the 2016 Nicaragua
438 epidemic. EBioMedicine. 2021 Oct 1;72.

439 24. Hozé N, Paireau J, Lapidus N, Kiem CT, Salje H, Severi G, et al. Monitoring the

proportion of the population infected by SARS-CoV-2 using age-stratified hospitalisation and serological data: a modelling study. Lancet Public Heal. 2021 Jun 1;6(6):e408–15.

25. Gordon A, Gresh L, Ojeda S, Katzelnick LC, Sanchez N, Mercado JC, et al. Prior dengue virus infection and risk of Zika: A pediatric cohort in Nicaragua. von Seidlein L, editor. PLOS Med. 2019 Jan 22;16(1):e1002726.

TABLES

Table 1. Summary and descriptive statistics of infection and disease outcomes across four epidemics in the PDCS and HICS in Managua, Nicaragua*

	First chikungunya epidemic (ChikE1)	Second chikungunya epidemic (ChikE2)	Zika epidemic (ZikaE)	COVID-19 epidemic (CovidE)
Cohort	PDCS	PDCS	PDCS	HICS
Participant age range	2-14	2-14	2-14	0-87
Epidemic period	9/2014 – 2/2015	7/2015 – 2/2016	1/2016 – 1/2017	3/2020 – 10/2020
Primary transmission pathway	*Aedes* mosquitoes	*Aedes* mosquitoes	*Aedes* mosquitoes	Respiratory droplets
Number at risk of infection	3,124	2,864	3,017	1,793
Number of infections (Number at risk of being a case)	199	710	1,416	1,039
Number of cases	90	416	494	306
Risk of infection (95% CI)†	6.4% (5.5%, 7.4%)	24.8% (23.2%, 26.6%)	47.1% (45.1%, 49.1%)	57.5% (54.1%, 60.9%)
Risk of disease (95% CI)†	45.6% (38.6%, 52.6%)	58.7% (54.9%, 62.4%)	35.4% (32.8%, 38.1%)	28.9% (25.5%, 32.5%)
Incidence rate (95% CI)†	2.9% (2.3%, 3.6%)	14.5% (13.2%, 16.0%)	16.6% (15.2%, 18.1%)	17.1% (14.8%, 19.6%)
Bias due to treating the incidence rate as the infection risk (%)‡	2.9% – 6.4% = –3.5%	14.5% – 24.8% = –10.3%	16.6% – 47.1% = –30.5%	17.1% – 57.5% = –40.4%
Bias due to treating the incidence rate as the disease risk (%)‡	2.9% – 45.6% = –42.7%	14.5% – 58.7% = –44.2%	16.6% – 35.4% = –18.8%	17.1% – 28.9% = –11.8%

ANOVA-based ICC for intra-household correlation of infection risk (95% CI)[§]	0.22 (0.17, 0.28)	0.21 (0.16, 0.27)	0.22 (0.17, 0.27)	0.30 (0.25, 0.35)
ANOVA-based ICC for intra-household correlation of disease risk (95% CI)[§]	0.14 (0.00, 0.51)	0.26 (0.09, 0.42)	0.28 (0.18, 0.37)	0.21 (0.15, 0.28)

[*]Abbreviations: ANOVA, analysis of variable; CI, confidence interval; GEE, generalized estimating equations; HICS, Household Influenza Cohort Study; ICC, intracluster correlation coefficient; PDCS, Pediatric Dengue Cohort Study

[†]GEE model estimates are presented.

[‡]Negative values indicate that the incidence rate underestimates the risk of infection and the risk of disease.

[§]Table S5 contains additional information.

458 **Table 2.** Sample size, infection, disease, incidence, and bias metrics from a serosurvey
459 augmented by cases identifiable by either active or passive surveillance for the 2016 Zika
460 epidemic in the Pediatric Dengue Cohort Study[*,†]

	Study design: Serosurvey and active case surveillance[‡]	Study design: Serosurvey and passive case surveillance[§]	Bias due to incomplete case ascertainment[¶]
Number at risk of infection	3,017	3,017	
Number of total infections (Number at risk of being a case)	1,416	1,416	
Number of cases (symptomatic infections)	494	133	
Number of subclinical infections	992	1,283	
Risk of infection (95% CI)[#]	47.1% (45.1%, 49.1%)	47.1% (45.1%, 49.1%)	47.1% – 47.1% = 0.0%[**]
Risk of disease (95% CI)[#]	35.4% (32.8%, 38.1%)	9.3% (8.0%, 11.0%)	9.3% – 35.4% = –26.1%
Incidence rate (95% CI)[#]	16.6% (15.2%, 18.1%)	4.4%[††] (3.7%, 5.2%)	4.4% – 16.6% = –12.2%
Bias due to treating the incidence rate as the infection risk (%)[¶]	16.6% – 47.1% = –30.5%	4.4% – 47.1% = –42.7%	–42.7% + 0.0% = –42.7%[‡‡] (= 4.4% – 47.1%)
Bias due to treating the incidence rate as the disease risk (%)[¶]	16.6% – 35.4% = –18.8%	4.4% – 9.3% = –4.9%	–4.9% + –26.1% = –31.0%[§§] (= 4.4% – 35.4%)

461

462 [*]Abbreviations: CI, confidence interval; GEE, generalized estimating equations; RT-PCR,
463 reverse transcription polymerase chain reaction

464 [†]Data in the bottom portion of this table represent the non-spatial version of Figure 6.

‡The first column is the full data for the Zika epidemic obtained by a serosurvey (to capture all infections) and active case surveillance (to capture all cases). Our active surveillance approach captured Zika cases with clinical profiles outside of standard Zika case definitions (8) and augmented RT-PCR with a serological algorithm built from five separate serological assay results (25). See the Appendix for more details.

§The second column includes the data collected by the serosurvey and Zika cases obtainable under passive surveillance (*e.g.*, using only standard Zika case definitions and RT-PCR). If a serosurvey had not been conducted, only the sample size (3,017) and the number of cases (133) would be known.

¶Negative values indicate that the incidence rate underestimates the risk of infection and the risk of disease, whether under active or passive case surveillance.

#GEE model estimates are presented.

**Results from a population-level serosurvey would not be impacted by active versus passive case surveillance at a health facility, so the risk of infection is the same under either active or passive case surveillance.

††Using passive case surveillance, as is standard, would result in this estimate of the incidence rate. This is the only metric estimable in the absence of a serosurvey.

‡‡The total bias due to treating the incidence rate, obtained using passively collected case data, as the true infection risk can be indirectly estimated by summing its constituent biases: the bias of treating the passive incidence rate as the passive infection risk (–42.7%) and the bias in the infection risk induced by incomplete case ascertainment (0.0%). A direct estimation of this compounded bias can also be achieved by subtracting the true infection risk (47.1%) from the

incidence rate based on passive surveillance data (4.4%). Without a serosurvey, it would not be possible to estimate the true infection risk and hence quantify the degree of bias. However, the lack of a serosurvey does not remove an existing bias. Thus, even without a serosurvey, –42.7% is the total bias that would result from inferring the true infection risk from an incidence rate based on passively collected case data.

[§§]The total bias due to treating the incidence rate, obtained using passively collected case data, as the true disease risk can be indirectly estimated by summing its constituent biases: the bias of treating the passive incidence rate as the passive disease risk (–4.9%) and the bias in the disease risk induced by incomplete case ascertainment (–26.1%). A direct estimation of this compounded bias can also be achieved by subtracting the true infection risk (35.4%) from the incidence rate based on passive surveillance data (4.4%). Without a serosurvey and active case surveillance, it would not be possible to estimate the true disease risk and hence quantify the degree of bias. However, the lack of a serosurvey does not remove an existing bias. Thus, even without a serosurvey, –31.0% is the total bias that would result from inferring the true disease risk from an incidence rate based on passively collected case data.

FIGURE LEGENDS

Figure 1. The neighborhoods of the study area in Managua, Nicaragua. The cemetery is shown in blue, and the study health center is indicated by a white triangle.

Figure 2. Epidemic curves for four epidemics in Managua, Nicaragua, on a weekly basis. Data for epidemics in the PDCS (A) and HICS (B) are shown. The duration of the annual sampling periods for serosurveillance of infection history is shown in green. The additional 2020 midyear sampling, instituted to capture the first COVID-19 wave, is shown in orange. The epidemic curves for the chikungunya and Zika epidemics reflect case counts that were confirmed by rRT-PCR and a serological algorithm, as detailed in the Appendix. Due to the retrospective collection of illness onset data from some HICS participants, the COVID-19 epidemic curve reflects 1) the date of acute sample collection from rRT-PCR-positive cases, 2) the date of illness onset as reported by ELISA-positive participants, or 3) a randomly selected date from the month in which ELISA-positive participants recalled experiencing illness consistent with COVID-19. The epidemic curves for the PDCS and HICS are purposefully shown in different panels as direct comparisons of case counts between cohorts of different sample sizes can result in misleading inferences.

Figure 3. Maps of the infection risk, disease risk, case-based incidence rate, and bias. The infection risk (A-D), disease risk (E-H), and incidence rate (I-L) across four epidemics are shown in one color palette, with warmer colors indicating higher values of the appropriate metric, and are set against a white background. The difference between infection risk and the incidence rate (bias induced by treating the incidence rate as the infection risk) (M-P) and the corresponding

526 bias for the disease risk (Q-T) are shown in another color palette. Bias panels, as they have a
527 different scale, are set against a gray background. Contour lines show changes in infection,
528 disease, and bias metrics corresponding to the scale bar of percentages to the right of each plot.
529 Maps were generated from generalized additive mixed models. A white triangle indicates the
530 study health center. Neighborhoods are outlined in gray. Columns in the figure correspond to the
531 chikungunya epidemics (2014, 2015), Zika epidemic (2016), and COVID-19 epidemic (2020) in
532 Managua, Nicaragua (left to right).

533

534 **Figure 4. Cluster detection analyses of the infection risk, disease risk, and case-based**
535 **incidence rate.** Clusters of infection risk (A-D), disease risk (E-H), and the incidence rate (I-L)
536 across four epidemics are shown. Panels depict the results of Kulldorf's spatial scan statistic
537 conducted in SaTScan. Hierarchical clusters are shown in dark colors; Gini clusters are shown in
538 light colors. Hierarchical clusters identify the most statistically likely clusters; Gini clusters
539 maximize outcome rates. Hotspots are shown in pink; coldspots are shown in blue. Cluster
540 centers are numerically labeled. Arrows show the kind of risk clusters that incidence clusters
541 resemble. A white triangle indicates the study health center. Neighborhoods are outlined in gray.
542 Columns in the figure correspond to the chikungunya epidemics (2014, 2015), Zika epidemic
543 (2016), and COVID-19 epidemic (2020) in Managua, Nicaragua (left to right). Table S2 contains
544 additional information for this analysis.

545

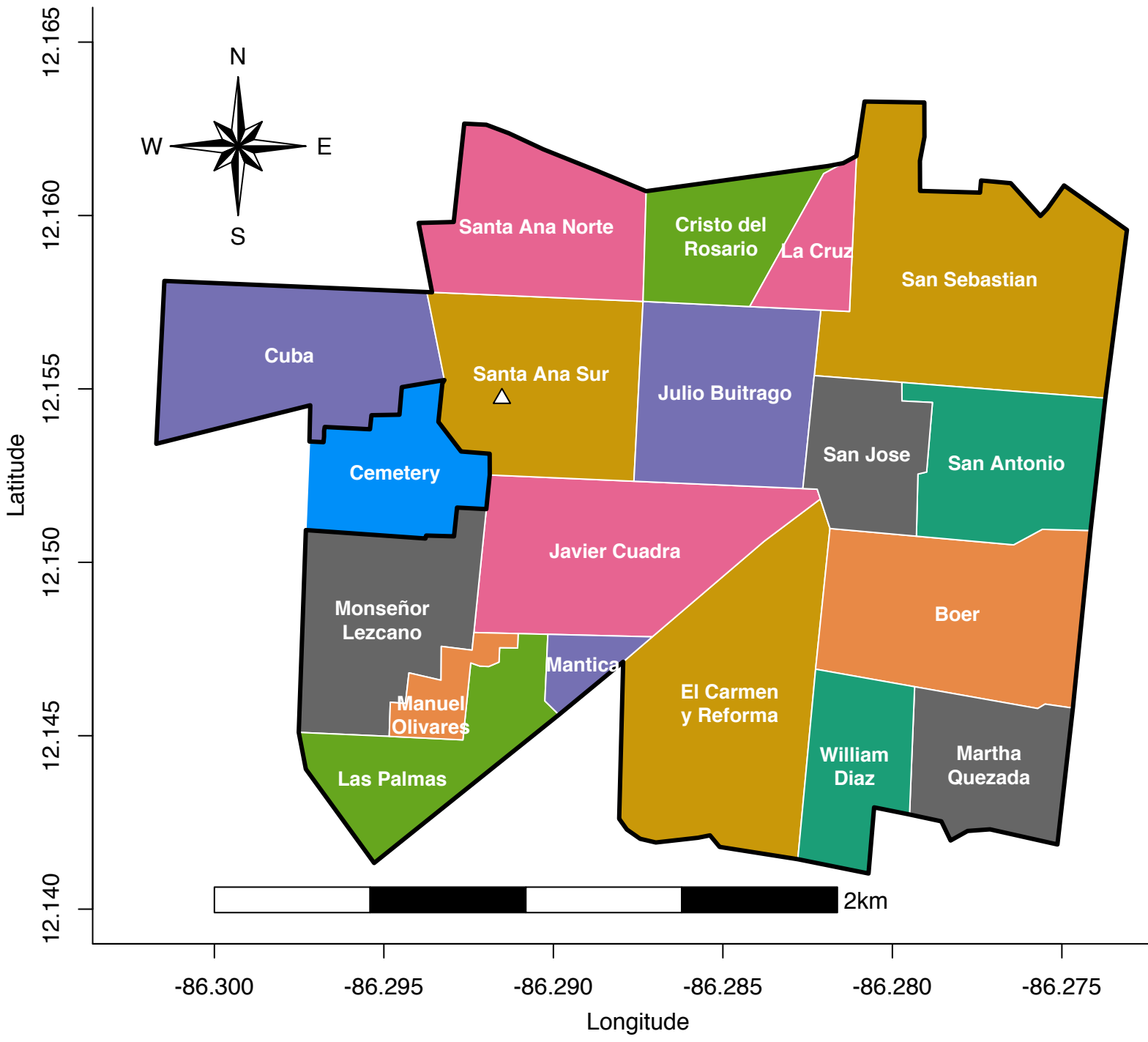
546 **Figure 5. Spatiotemporal dynamics across four epidemics in our study area.** Model
547 predictions of the infection risk (A-D, first column) and incidence rate (E-H, second column) are
548 reported per-month and per-1,000 population. Due to space constraints, data for months with few

549 cases are not shown. The PDCS epidemics (ChikE1, ChikE2, and ZikaE) are shown in a different
550 color palette than CovidE as the range of the infection dynamics for CovidE is so much higher
551 than that of the PDCS epidemics. Contour lines show changes in infection and incidence metrics
552 corresponding to the scale bar of percentages to the right of each plot. A white triangle indicates
553 the study health center. Neighborhoods are outlined in gray. Rows in the figure correspond to the
554 chikungunya epidemics (2014, 2015), Zika epidemic (2016), and COVID-19 epidemic (2020) in
555 Managua, Nicaragua (top to bottom).

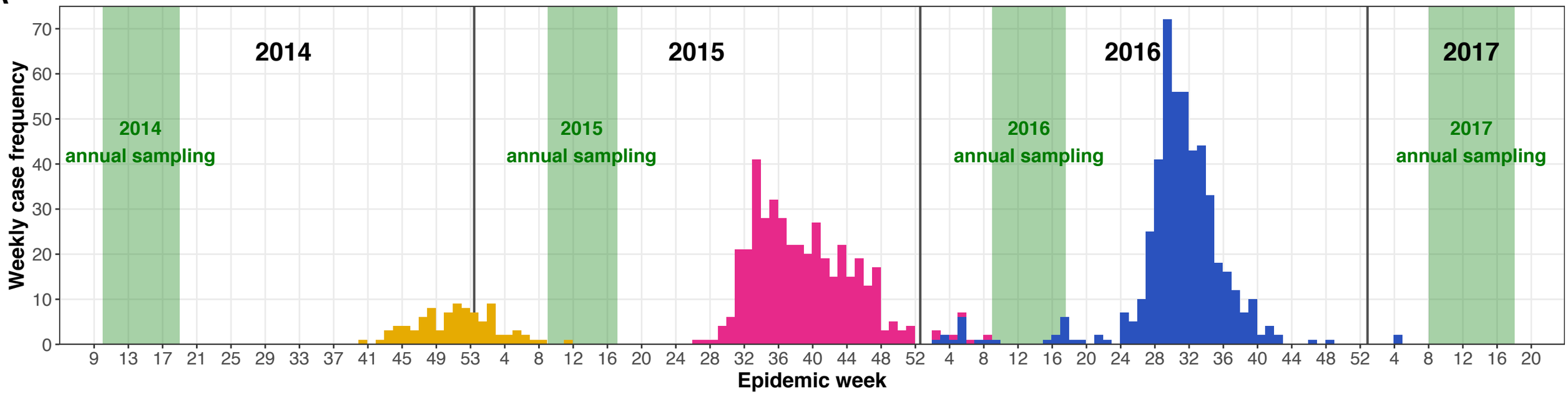
556

557 **Figure 6. Comparisons of infection, disease, incidence, and bias metrics for the 2016 Zika**
558 **epidemic in the PDCS using passive and active case surveillance.** Panels in this figure are
559 displayed in the same sequence as, and represent the spatial version of, data in the bottom portion
560 of Table 2. Columns in this figure correspond to a study design using a serosurvey and active
561 case surveillance (column 1), a study design using a serosurvey and passive case surveillance
562 (column 2), and the bias induced by passive versus active case surveillance (column 3). Column
563 1 repeats the data shown in Figure 3, column 3 for the sake of comparing the full data to that
564 obtained under passive case surveillance. Maps of the infection risk (A, F), disease risk (B, G),
565 and the incidence rate (C, H) are shown under active and passive case surveillance in the first
566 color palette and are distinguished by a white map background. The bias induced by active
567 versus passive case surveillance for these three metrics (K-M) is shown in a second color palette
568 distinguished by a green map background. The bias induced by treating the incidence rate as the
569 infection risk (D, I) and the incidence rate as the disease risk (E, J) is shown in a third color
570 palette distinguished by a grey map background. The total bias incurred from incomplete case
571 ascertainment and inferring a risk from the incidence rate (N-O) is shown in a fourth color

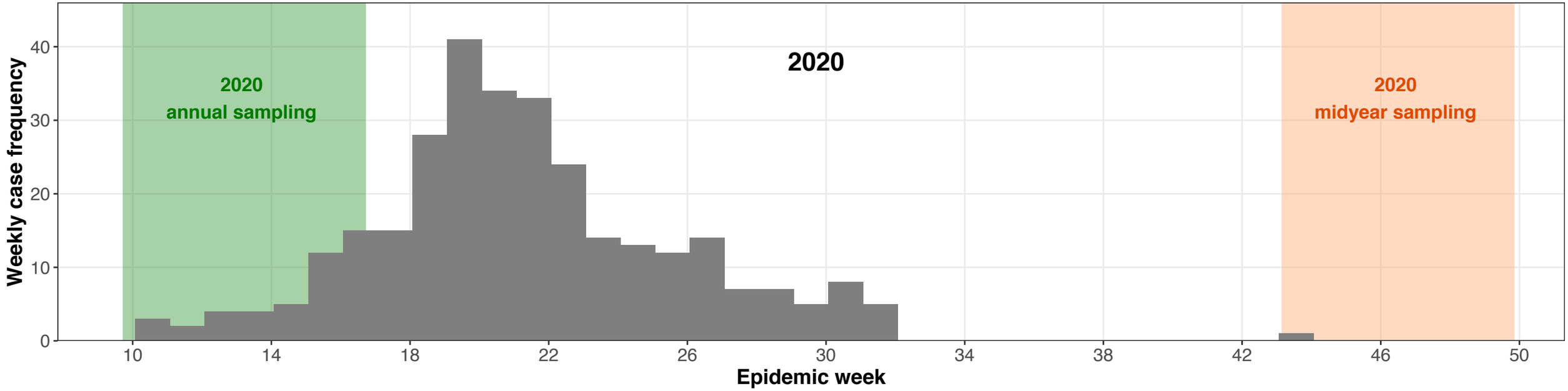
palette distinguished by a pink map background. Contour lines show changes in infection, disease, incidence, and bias metrics corresponding to the scale bar of percentages to the right of each plot. Maps were generated from generalized additive mixed models. A white triangle indicates the study health center. Neighborhoods are outlined in gray.

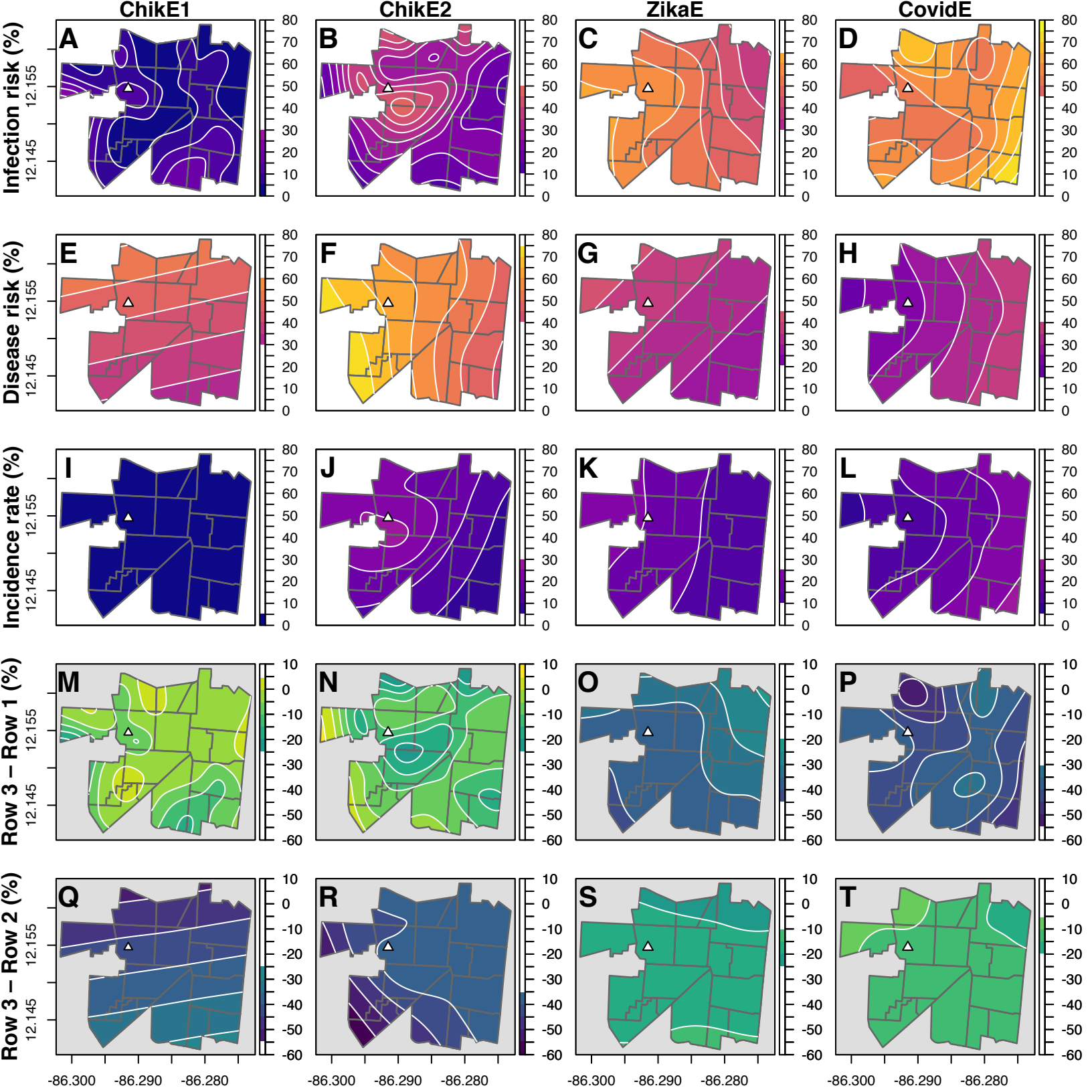

N
W
E
S
Santa Ana Norte
Cristo del Rosario
La Cruz
San Sebastian
Cuba
Santa Ana Sur
Julio Buitrago
San Jose
San Antonio
Cemetery
Javier Cuadra
Monseñor Lezcano
Boer
Mantica
Manuel Olivares
El Carmen y Reforma
William Diaz
Martha Quezada
Las Palmas
2km
12.165
12.160
12.155
12.150
12.145
12.140
Latitude
-86.300
-86.295
-86.290
-86.285
-86.280
-86.275
Longitude

A


2014
2015
2016
2017
2014
annual sampling
2015
annual sampling
2016
annual sampling
2017
annual sampling
Weekly case frequency
0
10
20
30
40
50
60
70
9 13 17 21 25 29 33 37 41 45 49 53 4 8 12 16 20 24 28 32 36 40 44 48 52 4 8 12 16 20 24 28 32 36 40 44 48 52 4 8 12 16 20
Epidemic week

B


2020
2020
annual sampling
2020
midyear sampling
Weekly case frequency
0
10
20
30
40
10 14 18 22 26 30 34 38 42 46 50
Epidemic week
PDCS Epidemics
Chikungunya 1
Chikungunya 2
Zika
HICS Epidemic
COVID-19


ChikE1
ChikE2
ZikaE
CovidE
Infection risk (%)
Disease risk (%)
Incidence rate (%)
Row 3 – Row 1 (%)
Row 3 – Row 2 (%)
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
12.155
12.145
-86.300
-86.290
-86.280
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60


Chikungunua epidemic 1
Chikungunua epidemic 2
Zika epidemic
COVID-19 epidemic
Infection risk clusters
Disease risk clusters
Incidence clusters
A
B
C
D
E
F
G
H
I
J
K
L
12.160
12.155
12.150
12.145
-86.300
-86.295
-86.290
-86.285
-86.280
-86.275