Latin American Data in Science, v.4, n.1
datainscience.com.br
ISSN: 2763-9290

Socio-demographic data of Zika records, Brazil, 2016-2021.

Recebido: 14/06/2024 | Aceito: 15/10/2024 | Publicado: 02/12/2024
https://doi.org/10.53805/lads.v4i1.66

Sebastião R. da S. Neto[1,2], Anna Beatriz Silva[1,2], Kayo Henrique de C. Monteiro[1,2], Élisson da S. Rocha[1,2], Thomás T. de Oliveira[1,2], Igor Vitor Teixeira[1,2], Raphael Augusto Dourado[1,2], Theo Lynn[3], Nguyen Tien Huy[4,5,6], Patricia T. Endo[*1,2]

ABSTRACT

The Zika virus (ZIKV) has emerged as a significant global health concern, particularly for pregnant women, given the potential complications it poses to the fetus. To effectively combat the disease, geospatial analysis of Zika cases has become increasingly important. By examining the incidence and distribution of Zika cases geographically, valuable insights can be gained, and high-risk areas can be identified. These findings are essential for formulating effective disease control measures. This article introduces a comprehensive Zika dataset based on data sourced from the Brazilian Notifiable Diseases Information System, encompassing the years 2016 to 2021. The dataset enables visualization and analysis of epidemiological information including case numbers, geographic distribution, spatial and temporal patterns, as well as common symptoms and complications associated with Zika infection.

Keywords: Zika; Data visualization; Data; Arbovirus.

[1] Universidade de Pernambuco (UPE), Programa de Pós-Graduação em Engenharia da Computação (PPGEC), Recife, Brazil. patricia.endo@upe.br
[2] dotLAB Brazil, Caruaru, Brazil.
[3] Dublin City University, Dublin, Ireland.
[4] Institute of Research and Development, Duy Tan University, Da Nang, Vietnam.
[5] School of Medicine and Pharmacy, Duy Tan University, Da Nang, Vietnam.
[6] School of Tropical Medicine and Global Health, Nagasaki, Japan.

DATA IMPORTANCE

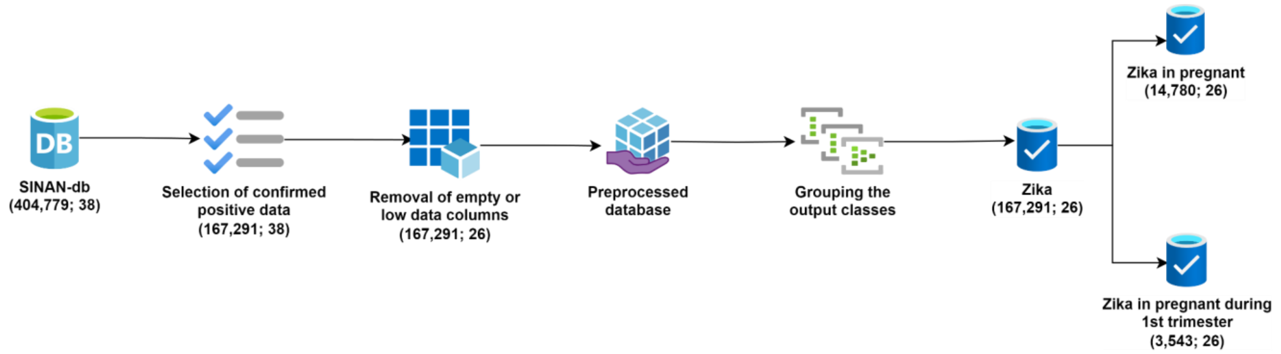
- Geospatial analysis of disease spread: The dataset allows the understanding of Zika virus cases (ZIKV), this helps to identify risk areas and trends in disease spread, which are crucial for effective resource distribution and strategic decision-making for disease control;
- Impact on vulnerable populations: The dataset includes filters for ZIKV cases in pregnant women and those in the first trimester, providing ideas about the impact on pregnancy;
- Sociodemographic analysis: The dataset highlights the influence of sociodemographic analysis factors on the spread of ZIKV. The study shows correlations between ZIKV incidence and low-income individuals, as well as areas with higher poverty rates and population density.

MATERIALS AND METHODS

The data was collected from SINAN, a public data repository that gathers information on the notified cases of diseases included in the national list of compulsory notifications. The focus of this study is on ZIKV cases reported in Brazil between 2016 and 2021. As discussed, while ZIKV was present in Brazil in 2015, universal mandatory notification was not required until February 2016 (WHO, 2022). The data covers all 26 Brazilian states and the Federal District of (Brasília). Unfortunately, the available Zika-related data is limited to sociodemographic information for each case, and no clinical or laboratory test details were provided. However, it is important to highlight that the available Zika-related data is limited to socio-demographic information for each case. No explanation was provided regarding the absence of data beyond clinical and laboratory tests.

The dataset initially contained 404,779 records of ZIKV cases with 38 attributes. To ensure data quality, a preprocessing step (see Figure 1) was performed, which involved removing incomplete or irrelevant attributes including filtering out cases that were not clinically confirmed. This resulted in a final dataset of 167,291 records and 26 attributes. To allow a more focused analysis, two additional filters were applied to the confirmed ZIKV cases: one for ZIKV in pregnant women and another for ZIKV in pregnant women during the first trimester. This decision was prompted by the significant impact of ZIKA incidence on pregnant women. This process identified 14,780 cases of ZIKV in pregnant women and 3,543 cases of ZIKV in pregnant women during the first trimester.

Figure 1.- Preprocessing steps performed to build the final dataset.


DB
SINAN-db (404,779; 38)
Selection of confirmed positive data (167,291; 38)
Removal of empty or low data columns (167,291; 26)
Preprocessed database
Grouping the output classes
Zika (167,291; 26)
Zika in pregnant (14,780; 26)
Zika in pregnant during 1st trimester (3,543; 26)

The dataset was originally composed of 404,779 records of Zika cases and 38 attributes collected from Brazilian Information System for Notifiable Diseases, Sistema de Informação de Agravo de Notifcação (SINAN), from 2016 to 2021. However, before proceeding, we performed a preprocessing step to clean the dataset by removing incomplete or irrelevant attributes.

Additionally, we applied a filter to consider only clinically confirmed cases. As a result, the final dataset consisted of 167,291 records and 9 attributes. Table 1 lists the attributes that we retained, while Table 2 presents the attributes that were removed.

The data contains notifications of Zika cases occurring in Brazil, encompassing all 26 states and the Federal District (Brasília). The related data does not include clinical information such as pre-existing symptoms or comorbidities, or laboratory test results. It only includes socio-demographic data for each case.

Table 1.- Attributes retained after preprocessing.

Attribute	Description
ID_AGRAVO	Name and code of the reported disease according to ICD-10.
DT_NOTIFIC	Notification date.
ID_REGIONA	Health region where the notifying municipality is located.
DT_SIN_PRI	Date of onset of severe symptoms.
SEM_PRI	Weeks of the standardized epidemiological calendar.
ID_MN_RESI	Municipality of residence.
ID_RG_RESI	Health region where the municipality of residence is located.
ID_PAIS	Country of residence.
DT_INVEST	Date of case investigation start.
TPAUTOCTO	Is the case autochthonous to the residence?
COUFINF	State (probable source of infection).
COPAISINF	Country (probable source of infection).
DT_ENCERRA	Closing date.
NU_ANO	Year of notification.
NU_IDADE_N	Age of the patient.
CS_SEXO	Sex of the patient.
CS_GESTANT	Gestational age of the patient.
CS_RACA	Race of the patient.
CS_ESCOL_N	Education level of the patient.
CLASSI_FIN	Classification (TARGET).
CRITERIO	Confirmation/discard criteria.
SEM_NOT	Epidemiological week in which the case was reported.
SG_UF_NOT	Table with codes and abbreviations.
ID_MUNICIP	Municipality of notification.
SG_UF	State of residence.
EVOLUCAO	Case evolution.

Table 2.- Attributes removed after preprocessing.

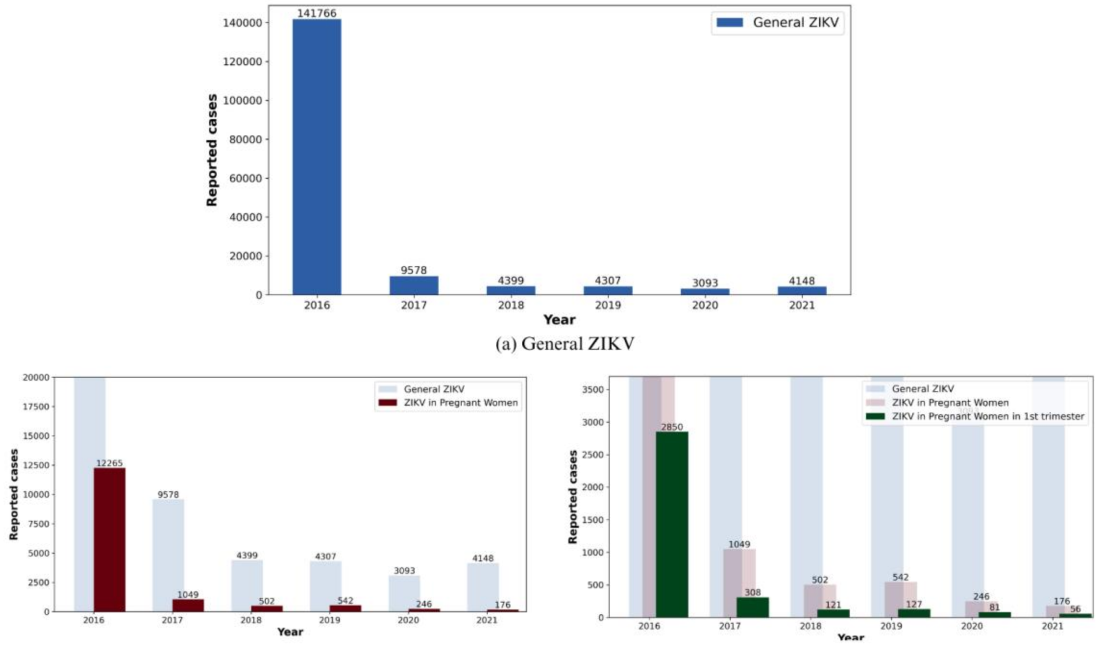
Attribute	Description
CS_SUSPEIT	Specifies the suspicion of the grievance
SEM_PRI	Weeks of the standardized epidemiological calendar
NDUPLIC_N	Not list/not included (logical removal)
IN_VINCULA	Notification binding with leprosy or tuberculosis
ID_OCUPA_N	Occupation of economic activity
COMUNINF	Municipality (probable source of infection)
DOENCA_TRA	Work-related illness
CS_FLXRET	Return stream
FLXRECEBI	Received by return stream (internal code)
TP_NOT	Notification type
TP_SISTEMA	No information in the dictionary

DATA DESCRIPTION

Dataset

The dataset, both in its processed and raw forms, is available in Mendeley Data at the suplementary materials section. Figure 2 presents the number of ZIKV records of ZIKV in the dataset categorized as General ZIKV, ZIKV in pregnant women, and ZIKV in pregnant women in the 1st trimester.

Figure 2.- Number of ZIKV records in the dataset by category (General ZIKV, ZIKV in Pregnant Women, and ZIKV in Pregnant Women in 1st trimester) in Brazil per year.



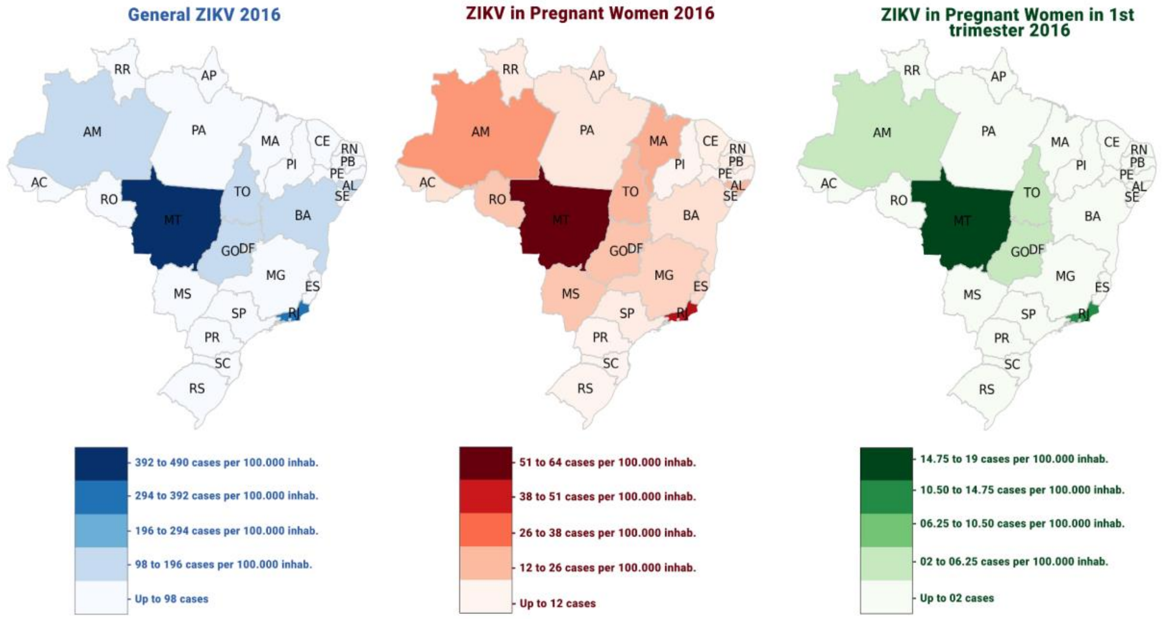
(a) General ZIKV

(b) ZIKV in Pregnant Women

(c) ZIKV in Pregnant Women in 1st trimester

The year 2016 witnessed and unprecedented surge in confirmed ZIKV cases compared to subsequent years. Figure 3 displays the states with the highest ZIKV incidence including categories such as General ZIKV, ZIKV in pregnant women, and ZIKV in pregnant women in the 1st Trimester. The figures and data presented here have been adjusted per 100.000 inhabitants to ensure accurate comparisons. In this representation, darker colors signify a higher number of reported cases. In the North region, Amazônia (AM) and Tocantins (TO) exhibited the highest incidence of ZIKV cases. Mato Grosso (MT) and Goiás (GO) were the standout states in the Midwest region. In the Northeast region, Bahia (BA) and Alagoas (AL) recorded notable numbers of ZIKV cases, adding to the regional burden of the disease. Lastly, among the Southeast region, Rio de Janeiro (RJ) had the highest number of Zika cases. The state's urban setting and population density might have contributed to the increased incidence.

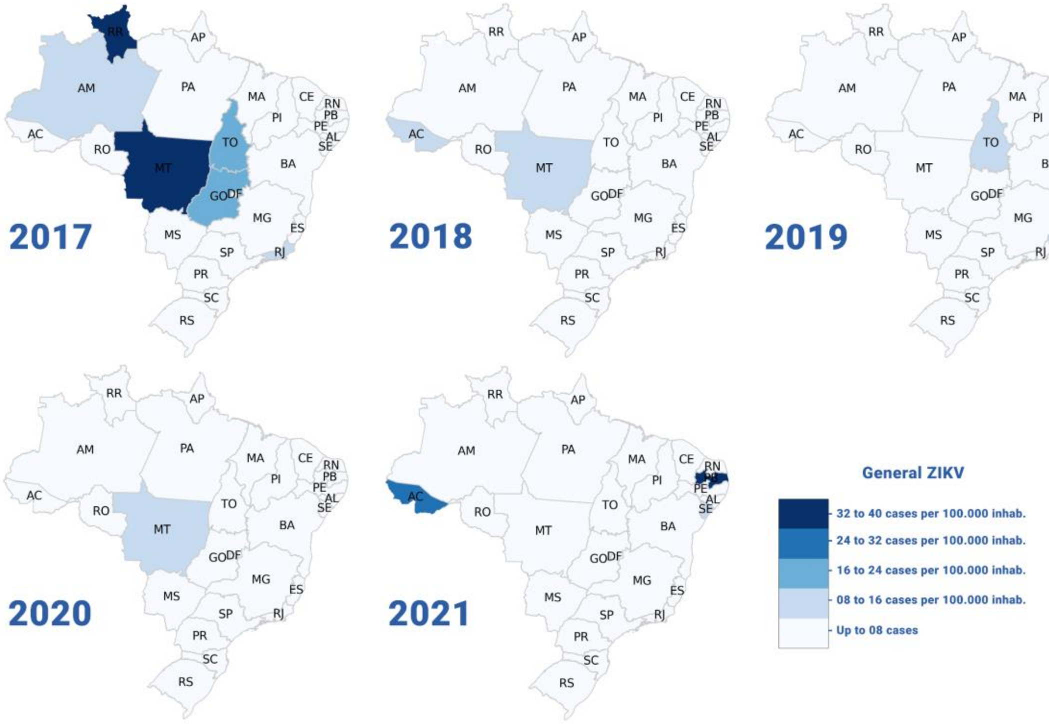
Figure 3.- Number of ZIKV records in the dataset by category (General ZIKV, ZIKV in Pregnant Women, and ZIKV in Pregnant Women in 1st Trimester) in Brazil by state in 2016.


General ZIKV 2016
RR
AP
AM
PA
MA
CE
RN
PB
PI
PE
AL
SE
AC
RO
TO
BA
MT
GO
DF
MG
ES
MS
SP
RJ
PR
SC
RS
392 to 490 cases per 100.000 inhab.
294 to 392 cases per 100.000 inhab.
196 to 294 cases per 100.000 inhab.
98 to 196 cases per 100.000 inhab.
Up to 98 cases
ZIKV in Pregnant Women 2016
RR
AP
AM
PA
MA
CE
RN
PB
PI
PE
AL
SE
AC
RO
TO
BA
MT
GO
DF
MG
ES
MS
SP
RJ
PR
SC
RS
51 to 64 cases per 100.000 inhab.
38 to 51 cases per 100.000 inhab.
26 to 38 cases per 100.000 inhab.
12 to 26 cases per 100.000 inhab.
Up to 12 cases
ZIKV in Pregnant Women in 1st trimester 2016
RR
AP
AM
PA
MA
CE
RN
PB
PI
PE
AL
SE
AC
RO
TO
BA
MT
GO
DF
MG
ES
MS
SP
RJ
PR
SC
RS
14.75 to 19 cases per 100.000 inhab.
10.50 to 14.75 cases per 100.000 inhab.
06.25 to 10.50 cases per 100.000 inhab.
02 to 06.25 cases per 100.000 inhab.
Up to 02 cases

Upon comparing the three presented filters across the year 2016, some patterns emerge. For instance, the state of MT had the highest rates per 100,000 inhabitants across all scenarios, followed by RJ, ranked second in overall ZIKV cases but led in absolute numbers within the Southeast region. Some states exhibited some uniqueness. Maranhão (MA), AL, Rondonia (RO), Mato Grosso do Sul (MS), and MG exhibited stronger shades for Zika in pregnant, indicating higher rates of ZIKV cases in pregnant women. Notably, when observing ZIKV cases for pregnant women in the 1st trimester, these states returned to lighter shades, suggesting a possible decline in incidence.

Figure 4 presents heatmaps of the number of ZIKV cases overall per 100,000 inhabitants overall, organized by state and year. The majority of ZIKV cases occurred in the Central-West and North regions of the country, specifically in the states of MT, GO, TO, and RR. In 2017, MT and RR had the highest number of ZIKV cases per 100.000 inhabitants. The significant increase in the state of Roraima (RR) may be related to the rise in the region's rainfall index. This indicates that the increase in rainfall led to a proliferation of Aedes aegypti mosquitoes and virus transmission in the state during this period.

Figure 4.- Number of ZIKV records in the dataset by General ZIKV by state.


2017
2018
2019
2020
2021
General ZIKV
32 to 40 cases per 100.000 inhab.
24 to 32 cases per 100.000 inhab.
16 to 24 cases per 100.000 inhab.
08 to 16 cases per 100.000 inhab.
Up to 08 cases
RR
AP
AM
PA
MA
CE
RN
PB
PE
AL
SE
PI
AC
RO
TO
BA
MT
GO
DF
MG
ES
MS
SP
RJ
PR
SC
RS

Overall, there is a noticeable decrease in cases over the years in all states. However, in 2021, two states, Acre (AC) and Paraíba (PA), stand out for the number of cases. These states recorded the highest number of Zika cases in Brazil when compared to previous years. According to the State Department of Health, the increase in cases in the state of Paraíba (PB) may be attributed to a low notification rate by municipalities in 2020, caused by the COVID-19 pandemic. This resulted in a drastic reduction in prevention and control measures against mosquitoes.

Figure 5 displays the distribution of ZIKV cases in pregnant women across different states. In 2017, the highest incidence was reported in the North and Central-West regions, as well as parts of the Northeast, and the states of Espírito Santo (ES) and RJ. Moving on to 2018, the state of MG stood out once again, along with AL and the same states in the Southeast region. By 2019, there was a decrease in cases in the North and CentralWest regions, with concentrations only in AL in the Northeast, and ES and RJ in the Southeast. Subsequently, in 2020 and 2021, the number of cases continued to decline. It is worth noting that this decline might have been influenced by the COVID-19 pandemic, although it cannot be solely attributed to this factor.

Figure 6 presents the incidence of ZIKV cases in pregnant women during the first trimester over the years 2016-2021. The data reveals interesting trends and regional variations. Following the WHO notification of the disease in Brazil, the number of ZIKV cases saw a steady increase across 2016, with the year 2017 experiencing the highest number of reported cases. This upward trend was particularly evident in the Central-West region, with the state of MT recording the highest incidence. Notably, the North and Northeast regions also reported cases during this period. The states of AM, Roraima (RR), RO, Ceará (CE), PA, and MA, respectively, have the highest number of cases in these regions. Moving to 2018, there was a nationwide decrease in ZIKV cases, with the regions of the North, Central-West, Northeast, and Southeast still reporting cases, though in lower numbers. In 2019, the state of Amapá (AP) in the North region, along with AL in the Northeast and ES and RJ in the Southeast, stood out for having a notable incidence of ZIKV cases. The years 2020 and 2021 continued the downward trend in the incidence of IKV cases. However, it is important to consider that these years may have been influenced by the COVID-19 pandemic. In 2021, only the Northeast region, specifically the states of PE and Sergipe (SE), reported ZIKV cases in pregnant women.

Figure 5.- Number of ZIKV records in the dataset by ZIKV in Pregnant Women by state.

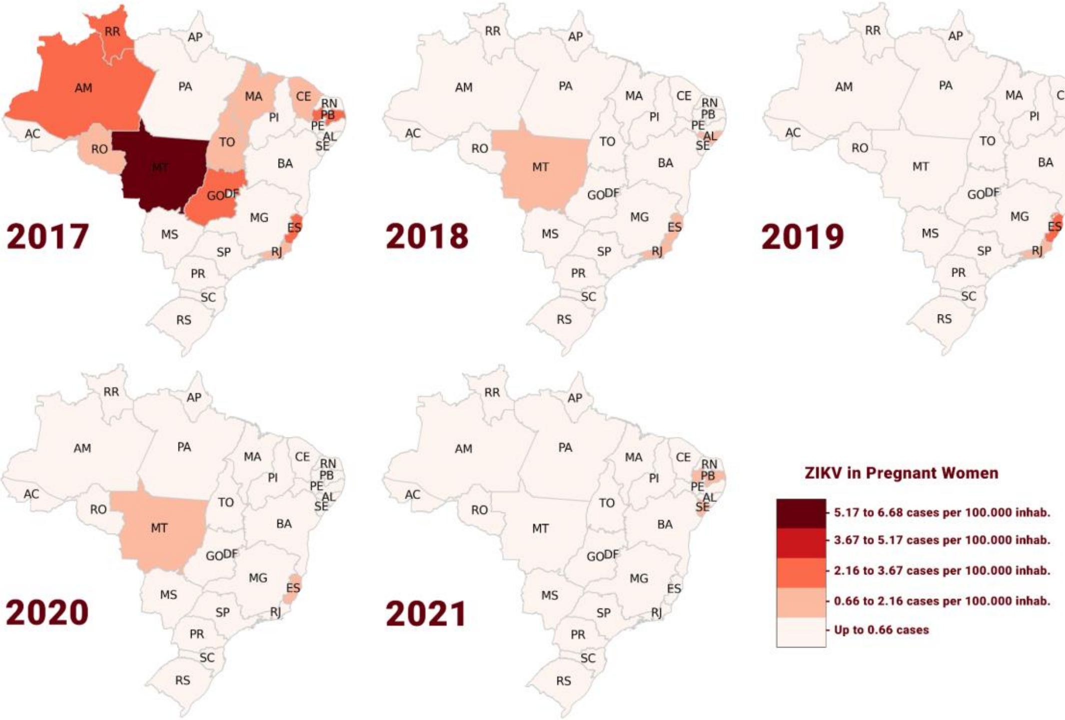

2017
2018
2019
2020
2021
RR
AP
AM
PA
MA
CE
RN
PB
PE
AL
SE
PI
AC
RO
MT
TO
BA
GO
DF
MG
ES
MS
SP
RJ
PR
SC
RS
ZIKV in Pregnant Women
5.17 to 6.68 cases per 100.000 inhab.
3.67 to 5.17 cases per 100.000 inhab.
2.16 to 3.67 cases per 100.000 inhab.
0.66 to 2.16 cases per 100.000 inhab.
Up to 0.66 cases

Figure 6.- Number of ZIKV records in the dataset by ZIKV in Pregnant Women by state.

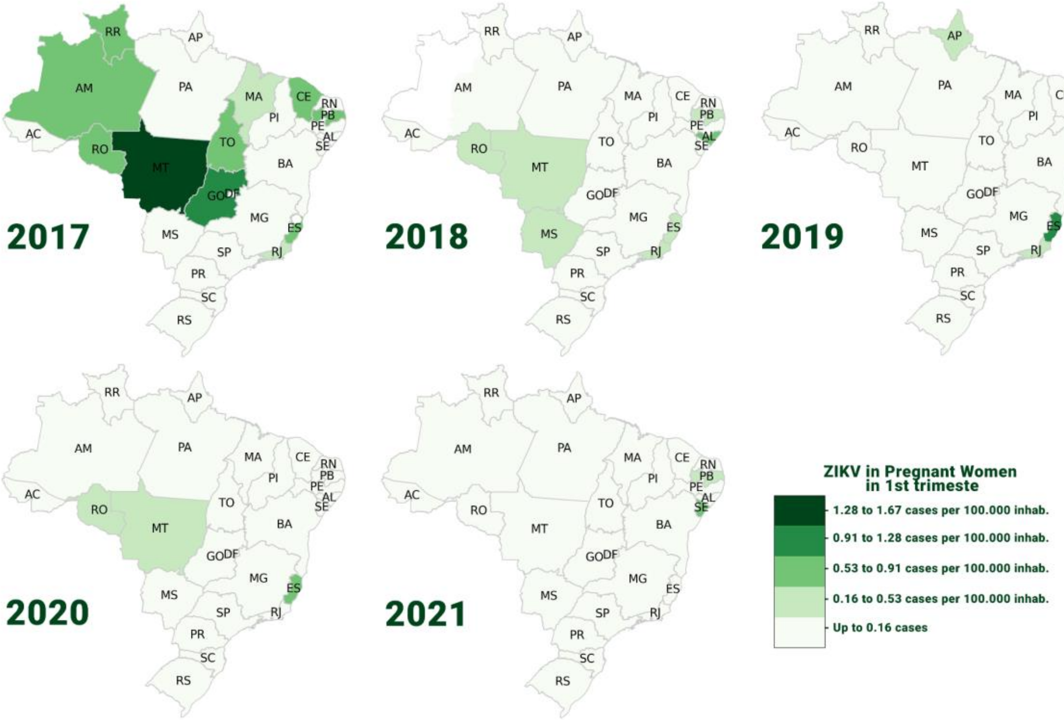

2017
2018
2019
2020
2021
RR
AP
AM
PA
MA
CE
RN
PB
PE
AL
SE
PI
AC
RO
MT
TO
BA
GO
DF
MG
ES
MS
SP
RJ
PR
SC
RS
ZIKV in Pregnant Women
in 1st trimeste
1.28 to 1.67 cases per 100.000 inhab.
0.91 to 1.28 cases per 100.000 inhab.
0.53 to 0.91 cases per 100.000 inhab.
0.16 to 0.53 cases per 100.000 inhab.
Up to 0.16 cases

Table 3 (in Anexs) presents various variables related to the ZIKV virus, providing important information about the affected patients. In total, 167,291 ZIKV cases were reported, with 14,780 occurring in pregnant women. Among pregnant women, 3,543 were affected during the 1st trimester of pregnancy. These numbers highlight the importance of investigating the impacts of ZIKV on pregnancy.

When analyzing the distribution by year, it is evident that the majority of cases occurred in 2016, representing 84.7% of total ZIKV cases, 83% of cases in pregnant women, and 80.4% of cases in pregnant women in the first trimester. However, in the following years, there was a gradual reduction in the number of cases, indicating a possible decrease in virus incidence.

The average age of affected patients is a relevant aspect to consider when discussing ZIKV. Generally, the average age for ZIKV cases is 33 years. However, for pregnant women with ZIKV, particularly those in the first trimester, the average age is lower around 26 to 27 years. This information is essential as it helps us identify the age group that is most affected by the virus.

Looking at the distribution by gender, it becomes evident that the majority of ZIKV cases occurred in females, accounting for 67.3% of the total cases. Upon analyzing the distribution of cases according to gestational age, it becomes apparent that the majority of reported instances fall into category 6 (Not applicable), accounting for 45.9% of cases. Following this, category 5 (No), represents the second-highest number of cases, making up 32.1% of the total. Together with the gender data, these findings imply a substantial portion of cases are related to male individuals. In cases involving women, most of them were not in a gestational period, indicating that Zika virus infection did not occur during pregnancy. It is essential to emphasize that these preliminary observations are solely based on the data available in the table. However, they could provide insights for future scientific investigations into risk factors and virus transmission.

Analysis of race distribution in ZIKV cases indicates that two races, race 1.0 (White) and race 4.0 (Mixed), are the most prevalent, accounting for 20.7% and 33.5% of cases, respectively. These findings raise the possibility of associations between and race and susceptibility to the virus. However, to gain a deep understanding of these relationships, further studies are necessary.

Regarding the education variable, it shows that the highest proportion of cases (41.1%) are unknown (education level 9.0). These findings might suggest a potential link between education level and ZIKV incidence. Nonetheless, it is essential to consider that the information about education level was unknown in a significant portion of cases thereby limiting the interpretation of these results. Additional research is needed to explore this relationship more comprehensively.

The “CRITERIO” column provides the diagnostic criteria used to identify ZIKV cases. It is worth noting that the majority of cases (88.3%) were diagnosed using criterion 2.0 (Clinical-epidemiological), while criterion 1.0 (Laboratory) was applied in 11.3% of cases. This information is esseDntial for evaluating the validity and reliability of the diagnostic criteria employed, as it highlights the predominance of clinical epidemiological diagnosis in identifying ZIKV cases.

Table 3 offers an overview of key characteristics of patients affected by the ZIKV virus, particularly focusing on pregnant patients and separately those patients in the first trimester of pregnancy. By examining this table, one can gain insights into various aspects of ZIKV infection, including distribution of cases over different temporal periods, the average age of affected patients, gender distribution, gestational age, racial background, education level, and diagnostic criteria used. This information is of paramount importance as it helps in understanding the epidemiology of the Zika virus and plays a crucial role in devising and implementing appropriate prevention and control measures. By having a clear and organized presentation of these characteristics, researchers and public health officials can make informed decisions to effectively combat the spread of the virus and protect vulnerable populations, especially pregnant women and their unborn children.

This dataset offers valuable information for researchers, policymakers, and healthcare professionals to combat the ZIKV effectively. It allows for various analyses, such as understanding the temporal distribution of cases, investigating the impact on pregnancy, exploring demographic factors, and conducting geospatial analysis to identify high-risk areas. These insights can inform targeted interventions and public health strategies to protect vulnerable populations.

SUPPLEMENTARY MATERIALS

Dataset: https://data.mendeley.com/datasets/fd83m2yj7j/1

ACKNOWLEDGEMENTS

This work was partially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE), and Universidade de Pernambuco (UPE), an entity of the Government of the State of Pernambuco focused on the promotion of teaching, research and extension.

REFERENCES

WHO - WORLD HEALTH ORGANIZATION et al. Zika epidemiology update–February 2022. Access on: November 11, 2024.

ANEXS

Table 3.- General and disease baseline characteristics.

Variables	General ZIKV	ZIKV in Pregnant Women	ZIKV in Pregnant Women in 1st trimester
N	167291/167291 (100.0%)	14780/167291 (8.8%)	3543/167291 (2.1%)
NU_ANO:2016	141766/167291 (84.7%)	12265/14780 (83.0%)	2850/3543 (80.4%)
NU_ANO:2017	9578/167291 (5.7%)	1049/14780 (7.1%)	308/3543 (8.7%)
NU_ANO:2018	4399/167291 (2.6%)	502/14780 (3.4%)	121/3543 (3.4%)
NU_ANO:2019	4307/167291 (2.6%)	542/14780 (3.7%)	127/3543 (3.6%)
NU_ANO:2020	3093/167291 (1.8%)	246/14780 (1.7%)	81/3543 (2.3%)
NU_ANO:2021	4148/167291 (2.5%)	176/14780 (1.2%)	56/3543 (1.6%)
NU_IDADE_N	32.9 (18.1)	26.8 (7.3)	27.0 (7.0)
CS_SEXO	112522/167291 (67.3%)	14779/14780 (100.0%)	3542/3543 (100.0%)
CS_SEXO	290/167291 (0.2%)	1/14780 (0.0%)	1/3543 (0.0%)
CS_SEXO	54479/167291 (32.6%)	-	-
CS_GESTANT:1.0	3543/167291 (2.1%)	3543/14780 (24.0%)	3543/3543 (100.0%)
CS_GESTANT:2.0	5755/167291 (3.4%)	5755/14780 (38.9%)	-
CS_GESTANT:3.0	4955/167291 (3.0%)	4955/14780 (33.5%)	-
CS_GESTANT:4.0	527/167291 (0.3%)	527/14780 (3.6%)	-
CS_GESTANT:5.0	53719/167291 (32.1%)	-	-
CS_GESTANT:6.0	76739/167291 (45.9%)	-	-
CS_GESTANT:9.0	22046/167291 (13.2%)	-	-
CS_RACA:1.0	34710/167291 (20.7%)	4152/14780 (28.1%)	1047/3543 (29.6%)
CS_RACA:2.0	6205/167291 (3.7%)	883/14780 (6.0%)	230/3543 (6.5%)
CS_RACA:3.0	976/167291 (0.6%)	148/14780 (1.0%)	33/3543 (0.9%)
CS_RACA:4.0	56075/167291 (33.5%)	6191/14780 (41.9%)	1482/3543 (41.8%)
CS_RACA:5.0	457/167291 (0.3%)	25/14780 (0.2%)	10/3543 (0.3%)
CS_RACA:9.0	52172/167291 (31.2%)	2803/14780 (19.0%)	616/3543 (17.4%)
CS_ESCOL_N:0.0	593/167291 (0.4%)	13/14780 (0.1%)	5/3543 (0.1%)
CS_ESCOL_N:1.0	4998/167291 (3.0%)	149/14780 (1.0%)	43/3543 (1.2%)
CS_ESCOL_N:2.0	3111/167291 (1.9%)	150/14780 (1.0%)	36/3543 (1.0%)

Variables	General ZIKV	ZIKV in Pregnant Women	ZIKV in Pregnant Women in 1st trimester
CS_ESCOL_N:3.0	8376/167291 (5.0%)	791/14780 (5.4%)	184/3543 (5.2%)
CS_ESCOL_N:4.0	4484/167291 (2.7%)	538/14780 (3.6%)	129/3543 (3.6%)
CS_ESCOL_N:5.0	7063/167291 (4.2%)	1187/14780 (8.0%)	294/3543 (8.3%)
CS_ESCOL_N:6.0	17473/167291 (10.4%)	2938/14780 (19.9%)	701/3543 (19.8%)
CS_ESCOL_N:7.0	2759/167291 (1.6%)	430/14780 (2.9%)	116/3543 (3.3%)
CS_ESCOL_N:8.0	5939/167291 (3.6%)	781/14780 (5.3%)	236/3543 (6.7%)
CS_ESCOL_N:9.0	68774/167291 (41.1%)	5640/14780 (38.2%)	1310/3543 (37.0%)
CS_ESCOL_N:10.0	12169/167291 (7.3%)	-	-
CRITERIO:1.0	18959/167291 (11.3%)	6966/14780 (47.1%)	1687/3543 (47.6%)
CRITERIO:2.0	147755/167291 (88.3%)	7759/14780 (52.5%)	1844/3543 (52.0%)