Publisher: Sapienza Grupo Editorial
R. Santa Cruz, 2187, Vila Mariana
São Paulo, Brazil
editor@sapienzaeditorial.com

Sapienza
International Journal of Interdisciplinary Studies
ISSN: 2675-9780

Assessing the quality of dengue data in the Philippines using Newcomb-Benford law

Avaliando a qualidade dos dados de dengue nas Filipinas utilizando a Lei de Newcomb-Benford

Evaluación de la calidad de los datos de dengue en Filipinas utilizando la Ley de Newcomb-Benford

Samuel John Parreño
https://orcid.org/0000-0002-2230-8984
Assistant Professor. Department of Teacher Education, University of Mindanao Digos College (UMDC). MSc Applied Mathematics – U. of Southeastern Philippines. BA Education – U. of Mindanao.
samueljohnparr@gmail.com (correspondence)

ARTICLE HISTORY

Received: 11-06-2023
Revised Version: 30-08-2023
Accepted: 06-09-2023
Published: 15-09-2023

Manuscript type: Article

ARTICLE INFORMATIONS

Science-Metrix Classification (Domain):
Economic & Social Sciences
Main topic:
Statistical data quality analysis
Main practical implications:
The main implication is the need for improved data quality and integrity in the reporting of dengue cases to enhance disease surveillance and control efforts in the Philippines.
Originality/value:
It presents a pioneering assessment of the adherence of official dengue cases data in the Philippines to Benford's distribution, providing valuable insights into data quality and emphasizing the need for improved integrity in disease surveillance and control.

ABSTRACT

Accurate and reliable data are vital for effective disease surveillance and control. This study examined the application of the Newcomb-Benford Law (NBL) as a tool for assessing the quality of dengue cases data in the Philippines. Large-scale datasets from the Epidemiology and Disease Control Surveillance (EDCS) and Philippine Integrated Disease Surveillance and Response (PIDSR) reports were analyzed to determine if the observed leading digit distributions deviate significantly from the expected NBL distribution. The statistical tests employed include the chi-squared test, Mantissa Arc Test, Mean Absolute Deviation (MAD), and distortion factor. The results reveal notable deviations from the expected NBL distribution, particularly in digits 1, 3, 4, 6, and 8, indicating potential irregularities and inconsistencies in the reported data. Factors contributing to these deviations may include data manipulation, measurement errors, and sampling biases. Improving data quality and integrity is crucial to ensure accurate disease surveillance.

Keywords: Dengue, First digit law, Newcomb-Benford law, Philippines.

RESUMO

Dados precisos e confiáveis são vitais para uma vigilância e controle eficazes de doenças. Este estudo examinou a aplicação da Lei de Newcomb-Benford (NBL) como uma ferramenta para avaliar a qualidade dos dados de casos de dengue nas Filipinas. Conjuntos de dados em larga escala dos relatórios de Vigilância e Controle Epidemiológico de Doenças (EDCS) e Vigilância e Resposta Integrada de Doenças das Filipinas (PIDSR) foram analisados para determinar se as distribuições de dígitos líderes observadas desviam significativamente da distribuição esperada pela NBL. Os testes estatísticos utilizados incluem o teste qui-quadrado, Teste de Arco de Mantissa, Desvio Médio Absoluto (MAD) e fator de distorção. Os resultados revelam desvios notáveis da distribuição esperada pela NBL, especialmente nos dígitos 1, 3, 4, 6 e 8, indicando possíveis irregularidades e inconsistências nos dados relatados. Fatores que contribuem para esses desvios podem incluir manipulação de dados, erros de medição e vieses de amostragem. Melhorar a qualidade e integridade dos dados é crucial para garantir uma vigilância precisa de doenças.

Palavras-chave Dengue, Lei do primeiro dígito, Lei de Newcomb-Benford, Filipinas.

RESUMEN

Datos precisos y confiables son vitales para la vigilancia y control efectivos de enfermedades. Este estudio examinó la aplicación de la Ley de Newcomb-Benford (NBL) como una herramienta para evaluar la calidad de los datos de casos de dengue en Filipinas. Se analizaron conjuntos de datos a gran escala de los informes de Vigilancia y Control de Epidemiología de Enfermedades (EDCS) y de Vigilancia Integrada de Enfermedades y Respuesta (PIDSR) de Filipinas para determinar si las distribuciones de los dígitos principales observados se desvían significativamente de la distribución esperada de NBL. Las pruebas estadísticas empleadas incluyen la prueba de chi-cuadrado, la prueba del arco de mantisa, la desviación absoluta media (MAD) y el factor de distorsión. Los resultados revelan desviaciones notables de la distribución esperada de NBL, especialmente en los dígitos 1, 3, 4, 6 y 8, lo que indica posibles irregularidades e inconsistencias en los datos reportados. Los factores que contribuyen a estas desviaciones pueden incluir manipulación de datos, errores de medición y sesgos de muestreo. Mejorar la calidad e integridad de los datos es crucial para garantizar una vigilancia precisa de enfermedades.

Palabras clave: Dengue, ley del primer dígito, Ley de Newcomb-Benford, Filipinas.

INTRODUCTION

Accurate and reliable data play a vital role in effective disease surveillance, prevention, and control efforts. The timely collection and analysis of disease-specific data in the field of public health are critical for informed decision-making and resource allocation (Stansfield et al., 2006; Grubin et al., 2021). The findings derived from analyzing public health data, particularly surveillance data, not only can contribute to enhancing healthcare provision and advocating for patient well-being but also serve as a foundation for determining priorities and allocating healthcare resources effectively (Soucie, 2012). Additionally, disease surveillance data can be used to produce forecasts, enabling governments to take appropriate actions to curb the rate of infection growth. However, collecting reliable and valid data is essential for accurately predicting or modeling diseases (Kolias, 2022). Ensuring the quality and integrity of health data can be challenging, particularly in resource-constrained settings like the Philippines. As the country grapples with the burden of infectious diseases such as dengue (de los Reyes & Escaner IV, 2018), it becomes essential to assess the reliability of reported case data to inform evidence-based interventions and actions.

Newcomb-Benford law (NBL), also known as the first-digit law, offers a statistical method that has been successfully employed to detect irregularities or anomalies in large datasets across various domains, including finance, auditing, and forensic accounting (Tam Cho & Gaines, 2007). The law states that in many naturally occurring datasets, the leading digits of numbers are not uniformly distributed but instead follow a specific logarithmic pattern, with smaller digits occurring more frequently than larger ones (Newcomb, 1881; Benford, 1938). This intriguing phenomenon has found applications in detecting fraud, data manipulation, and identifying potential errors or inconsistencies in numeric datasets (Moreau, 2021).

Prior research across various disciplines has utilized the NBL to identify instances of fraudulent or manipulated data including the distances between galaxies and known stars (Alexopolous & Leontsinis, 2014), asteroid and exoplanetary data (Melita & Miraglia, 2021), quantum phase transitions (De & Sen, 2011), multiple choice exams (Hoppe, 2016), crime statistics (Hickman & Rice, 2010), traffic in internet (Arshadi & Jahangir, 2014), medical images (Sanches & Marques, 2006), electoral processes (Pericchi & Torres, 2011), daily religious activities (Mir, 2014), income tax evasion (Nigrini, 2012), and COVID-19 data (Kolias, 2022). Since human-generated pseudo-random numbers frequently deviate from its expected distribution, the utilization of NBL has been widespread in detecting manipulated data (Gauvrit et al., 2017) and manipulated scientific articles (Hüllemann et al., 2017).

In the context of disease surveillance, applying NBL can provide valuable insights into the quality and reliability of reported cases. By comparing the observed distribution of leading digits in disease surveillance data against the expected distribution predicted by NBL, it is possible to identify potential anomalies, inconsistencies, or data manipulation that may indicate issues with data accuracy, reporting, or recording practices (Gauvrit et al., 2017). Such assessments can be particularly useful in low-resource settings, where limited resources and infrastructure may compromise the quality and reliability of disease data.

This paper aimed to explore the application of Newcomb-Benford Law (NBL) as a tool for assessing the quality of dengue cases data in the Philippines. Large-scale datasets obtained from national disease surveillance systems were analyzed and investigated whether the observed leading digit distributions deviate significantly from what is expected under NBL. By leveraging this statistical approach, we aimed to contribute to the ongoing efforts to enhance the accuracy and reliability of disease data and ultimately support evidence-based decision-making for improved public health outcomes.

The findings of this study have implications for public health policymakers, epidemiologists, and healthcare practitioners involved in disease surveillance and control. By identifying areas of concern and suggesting improvements in data quality, this research can contribute to enhancing the accuracy and reliability of disease data, ultimately leading to more effective prevention and control strategies.

METHODS

Data source

The data used in this study were secondary data retrieved from the Epidemiology and Disease Control Surveillance (EDCS) and Philippine Integrated Disease Surveillance and Response (PIDSR) weekly surveillance reports. These reports were published by the Philippine Department of Health (DOH) through the Public Health Surveillance Division of the Epidemiology Bureau. The EDCS and PIDSR reports encompass a comprehensive collection of disease-specific data, including information on dengue cases reported from various health facilities and local health offices across the country.

For the purpose of the analysis, data obtained from weeks 1 to 52 of 2019, weeks 1 to 10 of 2020, weeks 1 to 52 of 2021, weeks 1 to 52 of 2022, and weeks 1 to 24 of 2023 were used. The weeks excluded from the analysis were those in which no reports were available. These periods allowed us to capture a substantial timeframe, providing a robust dataset for assessing the quality of dengue cases data in the Philippines. By including data spanning multiple years, we aimed to account

for any temporal variations and identify potential long-term patterns or inconsistencies in the reported data.

Newcomb-Benford law

The NBL postulates that in naturally occurring datasets, the occurrence of the first digit follows a logarithmic pattern, with smaller digits (1, 2, 3) appearing more frequently than larger digits (7, 8, 9). Specifically, the distribution of first digits adheres to the following approximate proportions: 1 (30.1%), 2 (17.6%), 3 (12.5%), 4 (9.7%), 5 (7.9%), 6 (6.7%), 7 (5.8%), 8 (5.1%), and 9 (4.6%). Formally, the NBL is expressed by the following equation:

$$P(d) = \log_{10}\left(\frac{1+d}{d}\right) \text{ for } d \in \{1,2,\dots,9\}$$

where $P(d)$ is the probability of the digit d appearing (Silva & Figueiredo Filho, 2021).

In order to assess the degree to which the observed data align with the theoretical expectations of NBL, we employed the chi-squared (χ^2) test as a widely used statistical method for evaluating goodness of fit. Additionally, we employed the mantissa arc test, mean absolute deviation (MAD) and distortion factor (DF) as supplementary measures to ensure more robust and reliable results.

To conduct our analysis, we utilized R version 4.2.3. R offers a wide range of packages and functions for data analysis, including those specific to Benford's Law analysis. In particular, we employed the benford.analysis package developed by Cinelli to estimate the NBL functions and perform the necessary statistical tests (Cinelli, 2018).

RESULTS

For more accurate comparisons, we calculated formal tests to assess the goodness-of-fit. In this context, the null hypothesis assumes that the observed data adhere to the NBL. Hence, a smaller p-value indicates a greater level of confidence in rejecting the null hypothesis, suggesting that the observed distribution does not conform to the expected theoretical distribution of the NBL. As indicated in Table 1, the chi-squared test yields a value of 17.839 with 8 degrees of freedom. The associated p-value is less than 0.05, indicating a statistically significant deviation from the expected Newcomb-Benford distribution. The Mantissa Arc Test, which specifically examines the distribution of the fractional parts of the observed data, yields a value of 0.024 with 2 degrees of freedom. The corresponding p-value is less than 0.05, indicating a significant departure from the expected distribution.

Table 1. First digit distribution and test of significance of dengue cases data in the Philippines

Digit	NBL	Dengue Cases
1	30.1	33.9
2	17.6	20.1
3	12.5	15.3
4	9.7	5.8
5	7.9	6.4
6	6.7	2.1
7	5.8	3.7
8	5.1	8.5
9	4.6	4.2
N		189
χ^2		17.839*
MAD		0.028
DF		-6.178
Mantissa		0.024*

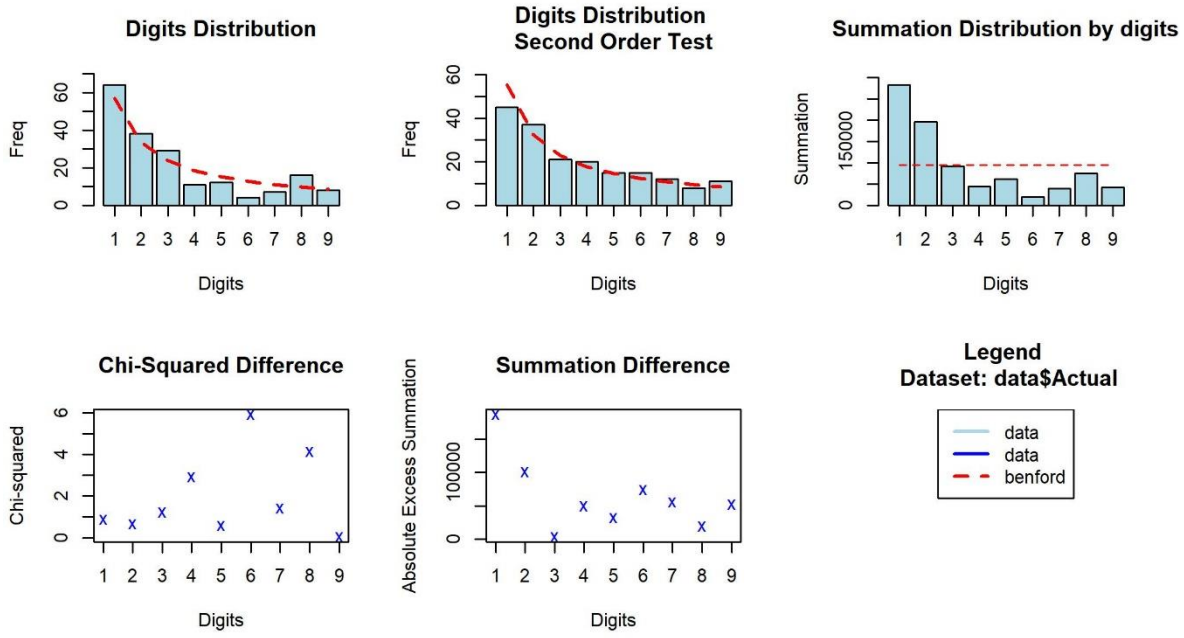
*p-value < 0.05

Source: Prepared by the author with research data

Like the z-statistic, the chi-squared test is susceptible to the influence of sample size, making it prone to rejecting the null hypothesis even for minor deviations from the expected distribution. On the other hand, the Mean Absolute Deviation (MAD) test is more robust as it does not take the number of records into account (Silva & Figueiredo Filho, 2021). The MAD

test evaluates the average difference between the observed and theoretical distributions, with a larger MAD indicating a greater average deviation. The MAD is calculated at 0.028 which is labeled as "nonconformity," indicating a notable deviation from the expected distribution (Nigrini, 2012). Finally, the Distortion Factor is reported as -6.178, suggesting a significant distortion in the observed data compared to the expected Benford distribution. Based on these criteria, it can be concluded that the dengue data from the Philippines significantly deviate from the expected distribution according to NBL. The five largest deviations from the expected Benford distribution are observed in the digits 1, 3, 4, 6, and 8. These digits have absolute differences of 8.65, 7.32, 7.11, 6.33, and 5.39, respectively. These deviations represent the largest deviations from the expected frequencies of these digits according to Benford's Law. (Figure 1).

Figure 1. NBL estimates of Dengue cases in the Philippines



Source: Prepared by the author with research data

DISCUSSION

The Newcomb-Benford Law (NBL) is a widely recognized statistical method used to identify potential anomalies in data. The application of NBL in finance and accounting often involves larger sample sizes compared to epidemiological data. This disparity in sample sizes may contribute to conflicting conclusions when applying different testing approaches (Nigrini, 2012; Koch and Okamura, 2020). Consequently, it is important to note that the results presented in this study do not serve as definitive evidence of suspicious activities. However, we have consistently observed significant deviations in the reported dengue cases data in the Philippines from the expected distribution according to NBL. These findings remain consistent across various empirical tests, strengthening our confidence in the conclusion that the current epidemiological surveillance system fails to provide reliable and trustworthy data on dengue cases in the Philippines.

When comparing our results with existing scientific articles that utilize the NBL, we find valuable insights and contextual understanding. In the study conducted by Gorenc (2019), which focused on detecting financial statement fraud using NBL, our findings align with theirs. They discovered significant deviations from the expected Newcomb-Benford distribution, indicating possible manipulation in the reported financial data. Similarly, our analysis identified notable departures from the expected distribution in the dengue cases dataset, supporting the presence of potential anomalies or irregularities. Another relevant study by Figueiredo Filho et al. (2022) examined the application of NBL in detecting data fraud in COVID-19 statistics. Their research revealed deviations from the expected distribution in certain countries' reported COVID-19 data, suggesting possible data irregularities. This finding resonates with our own results, which also identified significant deviations from the expected distribution in our dataset, implying potential anomalies or irregularities. Additionally, the study by Nigrini et al. (2007) explored the use of NBL in hydrology data analysis and highlighted its effectiveness in detecting anomalies and assessing data quality in that domain. Our findings align with their work, further supporting the notion that NBL is a valuable tool for assessing data quality in diverse fields, including healthcare and epidemiology.

In light of these results, it is important to consider the factors contributing to the deviations from the expected Newcomb-Benford distribution. Possible reasons could include data manipulation, measurement errors, sampling biases, or other systematic issues in data collection or reporting. Identifying the sources of these deviations is crucial to ensure the accuracy and reliability of the dataset. Furthermore, the findings highlight the need for additional measures to improve data quality and integrity. It may be necessary to reassess data collection processes, implement quality control measures, and enhance data validation procedures to minimize the occurrence of anomalies and improve the conformity of the data to NBL. Moreover, the statistical tests and metrics used in the analysis could be used as effective and cost-effective approach for assessing the adherence of the dengue cases data to NBL.

CONCLUSIONS

The application of Newcomb-Benford Law (NBL) as a tool for assessing the quality of dengue cases data in the Philippines has provided valuable insights. The analysis revealed significant deviations from the expected NBL distribution, indicating potential irregularities and inconsistencies in the reported data. The statistical tests, including the chi-squared test and the Mantissa Arc Test, confirmed these deviations, while the Mean Absolute Deviation (MAD) and distortion factor further highlighted the notable departure from the expected distribution. The five largest deviations were observed in the digits 1, 3, 4, 6, and 8, suggesting specific areas of concern.

The findings raised important considerations regarding the accuracy and reliability of disease surveillance data in the Philippines. Possible factors contributing to the deviations include data manipulation, measurement errors, and sampling biases. Addressing these issues is crucial to enhance data quality and integrity. It is recommended that data collection processes undergo reassessment, quality control measures be implemented, and data validation procedures be enhanced to minimize anomalies and improve conformity to the expected NBL distribution. The results of this study have implications for public health policymakers, epidemiologists, and healthcare practitioners involved in disease surveillance and control. By identifying areas of concern and suggesting improvements in data quality, this research contributes to enhancing the accuracy and reliability of disease data. Ultimately, these efforts will support evidence-based decision-making, leading to more effective prevention and control strategies.

Future research could focus on investigating the underlying causes of the observed deviations and exploring potential strategies to address them. Additionally, extending the analysis to other infectious diseases and expanding the dataset to include more recent years would provide a comprehensive assessment of data quality in disease surveillance. By continually evaluating and improving data quality, disease surveillance systems could be strengthened and public health interventions could be enhanced to better protect communities and promote well-being.

REFERENCES

Alexopoulos, T., & Leontsinis, S. (2014). Benford’s law in astronomy. Journal of Astrophysics and Astronomy, 35, 639-648.

Arshadi, L., & Jahangir, A. H. (2014). Benford's law behavior of Internet traffic. Journal of Network and Computer Applications, 40, 194-205.

Benford, F. (1938). The law of anomalous numbers. Proc Am Philos Soc, 78: , 551-572.

Cinelli, C. (2018). benford.analysis: Benford analysis for data validation and forensic analytics. Retrieved from https://github.com/carloscinelli/benford.analysis

de los Reyes, A. A., & Escaner IV, J. M. L. (2018). Dengue in the Philippines: model and analysis of parameters affecting transmission. Journal of biological dynamics, 12(1), 894-912.

De, A. S., & Sen, U. (2011). Benford's law detects quantum phase transitions similarly as earthquakes. Europhysics Letters, 95(5), 50008.

Figueiredo Filho, D., Silva, L., & Medeiros, H. (2022). "Won't get fooled again": statistical fault detection in COVID-19 Latin American data. Global Health, 18, 105. https://doi.org/10.1186/s12992-022-00899-1

Gauvrit, N. G., Houillon, J. C., & Delahaye, J. P. (2017). Generalized Benford’s Law as a lie detector. Advances in cognitive psychology, 13(2), 121.

Gorenc, M. (2019). Benford's Law As a Useful Tool to Determine Fraud in Financial Statements. Management (18544223), 14(1).

Grubin, L., Balachandran, L., Bartlett, S., Biritwum, N. K., Brooker, S., Fleming, F., Kollie, K., Matendechero, S., Mengistu, B., Muehleman, T. J., Mwingira, U., Partridge, B., Pavluck, A., Rebollo Polo, M., Tezembong, M., Treatman, D., Yeary, R., Zoerhoff, K., & Zoure, H. (2021). Improving data use for decision making by neglected tropical disease program teams: eight use cases. Gates open research, 5, 153. https://doi.org/10.12688/gatesopenres.13407.1

Hickman, M. J., & Rice, S. K. (2010). Digital analysis of crime statistics: Does crime conform to Benford’s law?. Journal of Quantitative Criminology, 26, 333-349.

Hoppe, F. M. (2016). Benford's law and distractors in multiple choice exams. International Journal of Mathematical Education in Science and Technology, 47(4), 606-612.

Hüllemann, S., Schüpfer, G., & Mauch, J. (2017). Application of Benford's law: a valuable tool for detecting scientific papers with fabricated data?: A case study using proven falsified articles against a comparison group. Der Anaesthesist, 66(10), 795-802.

Koch, C., & Okamura, K. (2020). Benford's law and COVID-19 reporting. Economics letters, 196, 109573.

Kolias, P. (2022). Applying Benford's law to COVID-19 data: the case of the European Union. Journal of Public Health, 44(2), e221-e226. https://doi.org/10.1093/pubmed/fdac005

Melita, M. D., & Miraglia, J. E. (2021). On the applicability of Benford law to exoplanetary and asteroid data. New Astronomy, 89, 101654.

Mir, T. A. (2014). The Benford law behavior of the religious activity data. Physica A: statistical mechanics and its Applications, 408, 1-9.

Moreau, V. H. (2021). Inconsistencies in countries COVID-19 data revealed by Benford's law. Model Assisted Statistics and Applications, 16(1), 73-79.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. Am J Math, 4: , 39.

Nigrini, M. J., & Miller, S. J. (2007). Benford's Law Applied to Hydrology Data—Results and Relevance to Other Geophysical Data. Mathematical Geology, 39, 469-490. https://doi.org/10.1007/s11004-007-9109-5

Nigrini, M. J. (2012). Benford's Law: Application for Forensic Accounting, Auditing and Fraud Detection. Wiley and Sons: New Jersey

Pericchi, L., & Torres, D. (2011). Quick anomaly detection by the Newcomb—Benford Law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. Statistical science, 502-516.

Sanches, J., & Marques, J. S. (2006, October). Image reconstruction using the Benford law. In 2006 International Conference on Image Processing (pp. 2029-2032). IEEE.

Silva, L., & Figueiredo Filho, D. (2021). Using Benford's law to assess the quality of COVID-19 register data in Brazil. Journal of public health, 43(1), 107-110.

Soucie, J. M. (2012). Public health surveillance and data collection: general principles and impact on hemophilia care. Hematology (Amsterdam, Netherlands), 17 Suppl 1(0 1), S144–S146. https://doi.org/10.1179/102453312X13336169156537

Stansfield, S. K., Walsh, J., Prata, N., et al. (2006). Information to Improve Decision Making for Health. In D. T. Jamison, J. G. Breman, A. R. Measham, et al. (Eds.), Disease Control Priorities in Developing Countries (2nd ed., Chapter 54). Washington, DC: The International Bank for Reconstruction and Development / The World Bank. Co-published by Oxford University Press, New York.

Tam Cho, W. K., & Gaines, B. J. (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. The american statistician, 61(3), 218-223

Contribution of each author to the manuscript:

	% of contribution of each author
Task	**A1**
A. theoretical and conceptual foundations and problematization:	100%
B. data research and statistical analysis:	100%
C. elaboration of figures and tables:	100%
D. drafting, reviewing and writing of the text:	100%
E. selection of bibliographical references	100%
F. Other (please indicate)	-

Indication of conflict of interest:

There is no conflict of interest

Source of funding

There is no source of funding

Acknowledgments

There is no acknowledgments