

12-2019

INSURING THE INTEGRITY OF GLOBAL REGISTRIES

AARDHRA MEENAKSHI VENKATACHALAM

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen



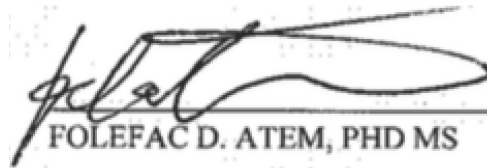
Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

INSURING THE INTEGRITY OF GLOBAL REGISTRIES

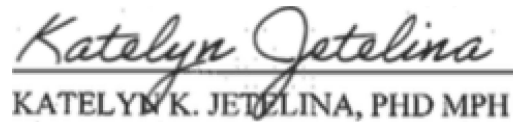
by

AARDHRA MEENAKSHI VENKATACHALAM, BA

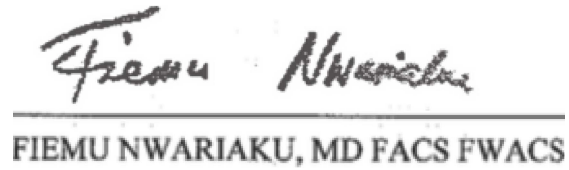
APPROVED:



FOLEFAC D. ATEM, PHD MS



KATELYN K. JETELINA, PHD MPH



FIEMU NWARIAKU, MD FACS FWACS

Copyright
by
Aardhra M. Venkatachalam, BA MPH
2019

INSURING THE INTEGRITY OF GLOBAL REGISTRIES

by

AARDHRA MEENAKSHI VENKATACHALAM
BA, Austin College, 2017

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF PUBLIC HEALTH

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
December 2019

ACKNOWLEDGEMENTS

Firstly, I would like to thank my family and friends for supporting me through my thesis. I would like to thank the Office of Global Health for providing insight into the data used for this thesis: Chinmayee Venkatraman, Chenchita Malohan and Erica Asante provided background on the dataset as well as coordinated meetings and assisted with data queries. A special thank you to my committee members, Dr. Folefac Atem, Dr. Katelyn Jetelina and Dr. Fiemu Nwaraiku for their guidance, as well as the UHealth School of Public Health for giving me this opportunity.

INSURING THE INTEGRITY OF GLOBAL REGISTRIES

Aardhra M. Venkatachalam, BA, MPH
The University of Texas
School of Public Health, 2019

Thesis Chair: Folefac Desiree Atem, MS PhD

Data quality is a priority in public health because it is used for public health program planning and implementation, research, as well as policy. The information gleaned from the data allows us to assess health status, and improve health needs. This paper evaluated methods for cleaning, managing and addressing missing data in the Lagos State Ambulance Service registry. The techniques included screening, data organization, diagnostic phase, treatment phase and a missing data phase to holistically handle the data. The percentage of systemic errors was reduced through the use of the 5-step method. Missing data was informative for state of origin, local government areas, age, gender and the event of a road traffic accident. Variables that were tested for missingness included occupation information, systolic and diastolic blood pressure, respiratory rate, Glasgow Coma scores for eye, verbal and motor, and the outcome of the EMS call.

TABLE OF CONTENTS

List of Tables	i
List of Figures	ii
Background	3
Methods	7
Data Analysis	9
Results	7
Conclusion	10
References	13

LIST OF TABLES

Table 1. Description of variables and percent missingness for LASAMBUS dataset.....	19
Table 2. Little's Chi-squared Test for MCAR/MAR.....	21

LIST OF FIGURES

Figure 1. Components of Public Health Data Quality.....	15
Figure 2. Distribution of Intervention forms by State of Origin from December 2017 to May 2019 for LASAMBUS dataset.....	16
Figure 3. Distribution of Intervention forms by Local Government Area from December 2017 to May 2019 for LASAMBUS dataset.....	17
Figure 4. Missing Data Analysis Decision Tree.....	18

BACKGROUND

Quality of Data's Impact

The functions of public health, like assessing health status, health needs, policy development, and allocation of resources, are heavily determined by data, which provides insight into the distribution of biomedical, behavioral, socioeconomic and environmental risk factors in the population [1]. The quality of data, specifically, plays an essential role in research, implementation of policy, healthcare practice, program planning and reporting [2]. Data quality, as defined by the International Standards Organisation (ISO), is “the features and characteristics of an entity which allow it to satisfy stated and implied needs”. [3] In other words, quality data is characterized by the accuracy and completeness of the dataset. Public health practice uses data-driven decisions, and poor data quality leads to a deficit in the ability to provide meaningful statistics for determining the health status of a population [4]. Some of the factors which reduce the integrity of the data include a weak data collection tool, data entry errors, data cleaning errors and finally, large amounts of information which is missing or random/ systemic errors which detract from the data accurately representing the population of interest. Poor data entry is difficult to correct if not caught in time [5], but with missing data, there are statistical procedures that can be employed to help explain the type of missing mechanism [6].

Country Health Information Systems, a review on trends in global health informatics by the World Health Organization (WHO), showed that while there has been progress in the

general status of health information and health monitoring, there remains an increasing gap in strengthening of data sources in low to middle income countries. [7] A study by Makombe et al. (2008) assessed the quality of site reports on national antiretroviral therapy in Malawi. The authors found that while sites had complete data for case registration and outcomes, the data was incomplete for critical points such as reasons for placement, treatment regimen and treatment outcome. Thus, the Malawi Ministry of Health was not able to draw conclusions for accurate national monitoring and data-driven policy for scaling up antiretroviral therapy programs [8].

The negative effects of poor data quality have been widely reported, and developing countries are most affected by poor data quality [9]. In recent years, government healthcare programs and administration, even at local levels, have experienced hardships with poor medical/health record documentation, inaccurate coding/classification of health outcomes and lack of access to national surveillance data for morbidity [10]. For example, poor data quality found in Routine Health Information Systems (RHIS) in Benin was associated with a significant drop in level of responsibility among healthcare administrators, and a decreased level of work engagement and perceived availability of resources [11]. Another study which assessed data availability and completeness found that irregular data quality assessments contributed to gaps in documentation, which were affecting interpretation of vital signs and adverse events in a medical male circumcision program in Zimbabwe [12].

Due to these challenges, accurate and accessible health data plays a large role in the quality of healthcare services provided. Statistics generated from poor quality public health

data inaccurately displays the use of services, prevalence of a health outcome or even prevents research from being conducted in needed topics [1, 2, 10].

Components of Data Quality

Data quality is composed of three main components (Figure 1): 1) the personnel involved; 2) the process of handling the data; and, 3) the elements of the data itself. Previous literature on data quality has identified these components to be essential to data quality, but do not mention how each component interacts with the other [13]. When these components complement each other, the integrity of the data increases through precision, relevance and generalizability. Precision refers to the relationship between the personnel and the process. Each individual contributing to specific parts of the process increases precision. Relevance refers to the relationship between data and the personnel. In other words, the personnel use their expertise to make sure that the data collected is biologically relevant to the exposure and outcome of interest. Generalizability refers to the relationship between the data and the data quality process. As data is cleaned and managed, it is essential to ensure that the process for maintaining data quality can be applied to new data added to the database/registry, and that the process is incorporates flexibility to changes made by the study team, or those who are working with the data.

Missing Data

Missing data is defined as the absence of values stored for a variable in an observation. [6] Missing data affects the internal and external validity of the analysis, and could lead to bias to analysis or sacrifice in statistical power. There are 3 main types of missing data, detailed by Little & Rubin [6]:

- 1) Missing Completely at Random (MCAR): means that missing data is not dependent on observed or unobserved values.
- 2) Missing at Random (MAR): given observed data, missingness does not depend on unobserved values.
- 3) Not missing at Random (NMAR): probability of the missing value depends on the values that are missing [14]

Figure 2 shows the Missing data analysis decision tree. For missingness less than 10%, traditional statistical techniques can still be used. For missingness greater than 10%, the type of missingness must be determined. NMAR is considered informative, as the outcome of interest is directly related to the missing data mechanism. MCAR and MAR are considered “non-informative”: while statistical power may be reduced, there is not an impact on estimates for the outcome of interest. For MCAR and MAR, the data is missing completely at random/without bias, or the missingness can be controlled. Informative/ Non-informative missingness is calculated using Little’s Chi-squared/T-test for MCAR. For large, significant quantities of informative missing data, the multiple imputation technique can be used to create substitute values based on the values observed in the registry. For large quantities of non-informative data, a complete case analysis or deletion technique can be used. For this project, imputation and case analysis were not used; the bivariate analyses were used as an indicator of the completeness of the registry.

Types of Data and Their implications

One type of data source commonly used are registries. Registries are large scale databases used by international organizations and local programs alike, and provide a source

for housing data. In recent years, developments in public health such as progress in information technology have made the use of registries more common. As with all sources of data, registries are subject to quality limitations such as missing values, bias, measurement error, computational error as well as human errors in data management⁴.

Although various data quality assessment mechanisms are available for the use and maintenance of registries, they have certain limitations⁵:

- Data quality checks are often carried out independently, even between different locations which contribute to the same registry. The process of data management is not standardized between locations.
- There is a lack in communication between the different departments/personnel who handle the data (Ex: Healthcare provider/Epidemiologist, Statistician, Data Manager and data collection personnel), which leads to inconsistencies in data management.
- Many large-scale registries/databases tend to have large quantities of missing data, which are accounted for, but not dealt with. Therefore, the impact of the missing data on the health outcome of interest is unknown.⁶ In some cases where data is obtained from multiple sources, a lack of standardization may lead to further data error.

Specific Aims

The aim of this study is two-fold:

- 1) To use a 5-step method modified from Van de Broeck et al. [14] to effectively clean and manage the Lagos State Ambulance Service (LASAMBUS) dataset;
- [2\)](#) Conduct a missing data analysis to assess if the missing data is informative or non-informative.

We hypothesize that the 5-step method will improve data quality by reducing systematic errors and missing data analysis will provide information on the next steps for handling the LASAMBUS dataset.

Public Health Significance

Public health practice is moving towards using more Big data, which means that the integrity of the data used determines how accurately the information collected represents true population values. With a standardized method for cleaning and managing data, and handling missing data, professionals from any scope of public health can insure that the data that they use for research, surveillance/monitoring or program implementation represents the true observations from the population of interest.

In traditional manuscripts, missing data is scarcely mentioned, but when used as a data quality tool, missing data analysis becomes a powerful statistical tool which can allow the public health professional to make more inference with data collected. Data collection and management are influenced by a variety of factors such as the type of data collection tool used, the amount of time the population of interest is observed, response rates and loss to follow up, some of which can be mitigated with information from missing data analysis.

METHODS

Study Population and Data Source

We used the Lagos State Ambulance Service (LASAMBUS) dataset. The LASAMBUS was constructed from intervention forms which asked about location (e.g., plate numbers of ambulance, state of origin for patients, Local Government area of EMS call), demographics (i.e., age, gender, occupation), vitals (e.g., Glasgow Como score, blood pressure readings, SP0₂, etc.) as well as details about the type of incident and the outcome of the emergency medical services (EMS) calls for service in Lagos, Nigeria. There were a total of 1,531 unique observations in the dataset, and data collection lasted from December 2017 to May 2018.

With the increase in urbanization, there has been an increase in the number of road traffic and industrial accidents, which supported the need for a prehospital system in Lagos, Nigeria. This need led to the establishment of the first emergency medical system in 2001. Before this time, the ambulances were known more as vehicles in undertaker services, and less as a resource to reduce the gap in access to clinical care [15]. The current Nigerian EMS system faces large quantities of trauma injury, from oil pipe explosions and collapsing structures, and has been weakened by the lack of federal funding [16].

The methods for this study followed five general phases:

1. **Screening:** Enrollment of participants, observation of inclusion/exclusion criteria.

The screening phase will describe the type of data in the Lagos Ambulance Dataset and how the data was collected. Since the data has already been collected by the

UTSW Office of Global Health, data quality checks for this phase will be determined by looking at de-identified data collection forms and recording inconsistencies in data entry/data collection.

2. **Data Organization:** Treatment of blank cells, inconsistencies in column variables, highlighting errors, data validation. This phase will treat the medium of data collection (excel spreadsheet). Basic data cleaning techniques will be used to ready the registry for analysis. Data quality checks for this phase will include making sure that the appropriate columns and rows have been consistently merged/appended.
3. **Diagnostic:** Check the interaction of the data with literature/clinical/epidemiological relevance. This phase checks that the data is biologically possible and make sense within the context of each variable. Data quality checks will include a data validation command run from the excel spreadsheet.
4. **Treatment:** Make necessary changes in spreadsheet and run descriptive statistics. Check interaction of variables with clinician/epidemiologist. In this phase, we reran the descriptive statistics and checked the output interaction to make sure that the numbers were relevant to the hypothesis.
5. **Missing Data:** We ran a Little's Chi-Squared/T-test for missing data to determine if the data was informative or non-informative. For any informative data, we reassessed the variables involved, the study design as well as anything else that could have contributed to the missing data.

Data Analysis

Microsoft excel was used to house the LASAMBUS dataset. Data validation techniques included [17]:

- 1) Format check: the data is in the correct format for different types of data
Ex: License plates numbers used in the LASAMBUS
- 2) Presence Check: the data has been entered into the correct field. Ex: Data was checked for row/column shifts, and corrected accordingly
- 3) Range check: values for each variable fell within an acceptable range. For vital signs and other physiological metrics what were recorded, outliers were checked with the intervention forms for data entry accuracy

Descriptive statistics were run for percent missingness for the variables of interest. Select variables (state, Local Government Area, age and gender) were recoded as either having data present or missing data, then compared to other variables to determine if data was missing completely at random (MCAR) or missing at random (MAR) using the Little's Chi-squared test for MCAR.

HUMAN SUBJECTS

This project was approved as a Quality Improvement (QI) Project by the University of Texas Health Science Center (UTHSC) Institutional Review Board (IRB). The LASAMBUS dataset creation and use was approved by the UT Southwestern Medical Center Institutional Review Board. The IRB approval number was STU 2018-8843 under the study titled "Lagos State Ambulance Service (LASAMBUS) Performance Evaluation".

RESULTS

Figure 2 and Figure 3 displayed the distributions of intervention reports filed based on the state of origin and local government areas, respectively. The state of origin reported on the intervention forms showed that Lagos state received large volumes of people from out of state. States such as Oyo, Osun, Ogun and Ondo are closer in proximity to Lagos state, which was reflected by the greater number of the intervention forms included in the registry. Other states such as Gombe and Yobe had few, if any intervention forms in the registry. (Figure 2). The Local Government Areas for Lagos state also showed distribution to be relatively even for the intervention forms (Figure 3). Despite the large amounts of missingness in State of Origin (~30%) and Local Government Area data (~26%), the information available for both showed a similar distribution. Data validation techniques were used to identify and correct the format, presence and range of the values resulted in a clean and analysis-ready dataset.

Table 1 displayed descriptive statistics for the variables in the LASAMBUS dataset.

Descriptive statistics suggest 7% missingness for plate number of ambulance, 69% for age, 39% for location from which call was made, 64% for gender, 73% for occupation information present, 76% for Local Government Areas, 73% for the three Glasgow Coma scores (GCS) (eye, verbal and motor) and 6% for the outcome of the EMS call.

Table 2 displayed the results of Little's Chi-squared/T-test for missingness for variables compared to indicator variables such as State of origin, local government area, age gender and an event of a road traffic accident. Occupation information was informative for state of origin ($p=0.006$) at an alpha of 0.05. Occupation information ($p<0.0001$), systolic blood pressure($p=0.036$) and diastolic blood pressure($p=0.031$) were informative for Local Government Area at an alpha of 0.05. Systolic blood pressure ($p<0.0001$), diastolic blood pressure ($p<0.0001$) and respiratory rate ($p=0.046$) were informative for age at an alpha of 0.05. All variables tested were non-informative for gender. All variables were informative for the event of a Road Traffic Accident ($p<0.001$) at an alpha of 0.05.

CONCLUSION

The outcome of EMS call played a large role in the type of missing data present in the LASAMBUS dataset. For example, if the outcome was that EMS responded to a false call or that the reason for the call was already addressed by before EMS responded, there was a higher chance that vitals such as SPO_2 , blood pressure and GCS scores were not recorded. Because there was no victim/patient. For future analyses, a subset of the data with only outcomes involving a victim/patient could reduce structural missingness, and more accurately demonstrate the prevalence of missing data in the registry. Limitations of the study included the effect of the outcome variable in the availability of the patient data. Specific outcomes for the EMS call would not have resulted in the EMS team attending to a patient, which means that the data on the patient would have most likely been recorded. Other outcomes, such as the "false call," would not have resulted in the EMS team attending

to a patient, so no data on the patient would have been collected. Other limitations included the fact that the intervention forms were paperwork required to obtain government funding. An intervention form which was created for public health use could have employed techniques to reduce missing data, as well as organize variable in a way that could have given more information about the population. The data management and missing data analysis allowed for these discrepancies to be identified so they can be addressed as the registry grows.

REFERENCES

1. Chen, H., et al., *A review of data quality assessment methods for public health information systems*. International journal of environmental research and public health, 2014. **11**(5): p. 5170-5207.
2. Cai, L. and Y. Zhu, *The challenges of data quality and data quality assessment in the big data era*. Data science journal, 2015. **14**.
3. Arts, D.G., N.F. De Keizer, and G.-J. Scheffer, *Defining and improving data quality in medical registries: a literature review, case study, and generic framework*. Journal of the American Medical Informatics Association, 2002. **9**(6): p. 600-611.
4. Keller, S., et al., *The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches*. Annual Review of Statistics and Its Application, 2017. **4**: p. 85-108.
5. Barchard, K.A. and L.A. Pace, *Preventing human error: The impact of data entry methods on data accuracy and statistical results*. Computers in Human Behavior, 2011. **27**(5): p. 1834-1839.
6. Little, R.J. and D.B. Rubin, *Statistical analysis with missing data*. Vol. 793. 2019: John Wiley & Sons.
7. Organization, W.H., *Country health information systems: a review of the current situation and trends*. Geneva: WHO, 2011.
8. Makombe, S.D., et al., *Assessing the quality of data aggregated by antiretroviral treatment clinics in Malawi*. Bulletin of the World Health Organization, 2008. **86**: p. 310-314.
9. Organization, W.H., *Improving data quality: a guide for developing countries*. 2003, Manila: WHO Regional Office for the Western Pacific.
10. Cheng, P., et al., *The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding*. Health Information Management Journal, 2009. **38**(1): p. 35-46.
11. Ahanhanzo, Y.G., et al., *Factors associated with data quality in the routine health information system of Benin*. Archives of public health, 2014. **72**(1): p. 25.
12. Xiao, Y., et al., *Challenges in data quality: the influence of data quality assessments on data availability and completeness in a voluntary medical male circumcision programme in Zimbabwe*. BMJ open, 2017. **7**(1): p. e013562.
13. Lee, Y.W., et al., *Journey to data quality*. 2009: The MIT Press.
14. Van den Broeck, J., et al., *Data cleaning: detecting, diagnosing, and editing data abnormalities*. PLoS medicine, 2005. **2**(10): p. e267.
15. Ibrahim, N.A., et al., *Road traffic injury in Lagos, Nigeria: assessing prehospital care*. Prehospital and disaster medicine, 2017. **32**(4): p. 424-430.
16. Adeloye, D., *Prehospital trauma care systems: potential role toward reducing morbidities and mortalities from road traffic injuries in Nigeria*. Prehospital and disaster medicine, 2012. **27**(6): p. 536-542.

17. Elliott, A.C., et al., *Preparing data for analysis using Microsoft Excel*. Journal of investigative medicine, 2006. **54**(6): p. 334-341.

Figure 1. Components of Public Health Data Quality

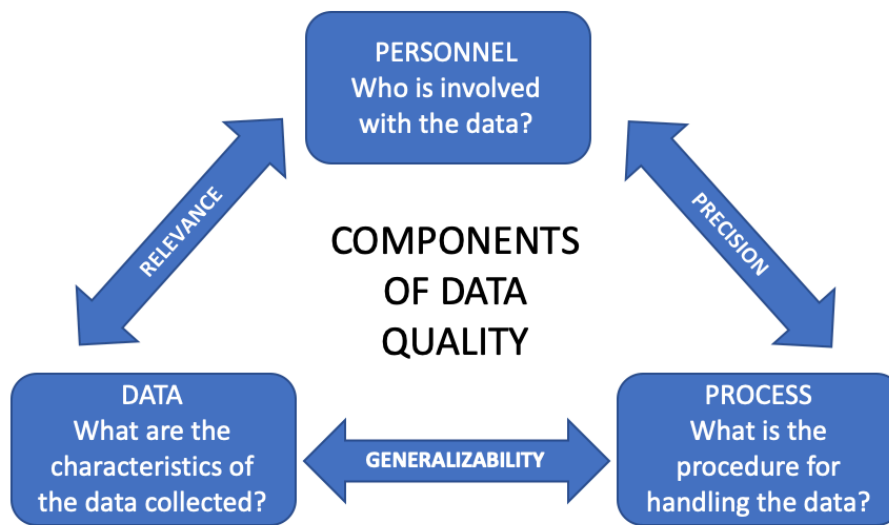


Figure 2. Distribution of Intervention forms by State of Origin from December 2017 to May 2019 for LASAMBUS dataset

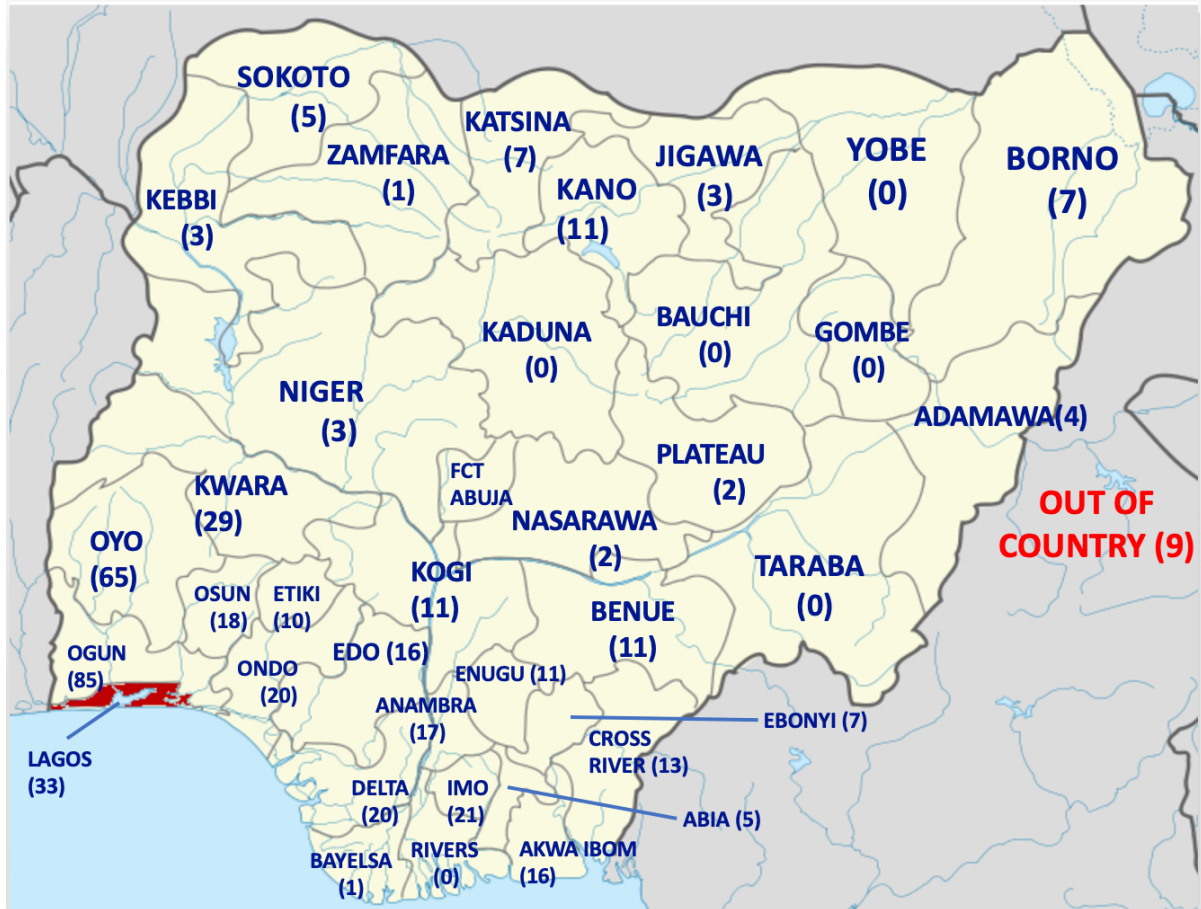


Figure 3. Distribution of Intervention forms by Local Government Area from December 2017 to May 2019 for LASAMBUS dataset

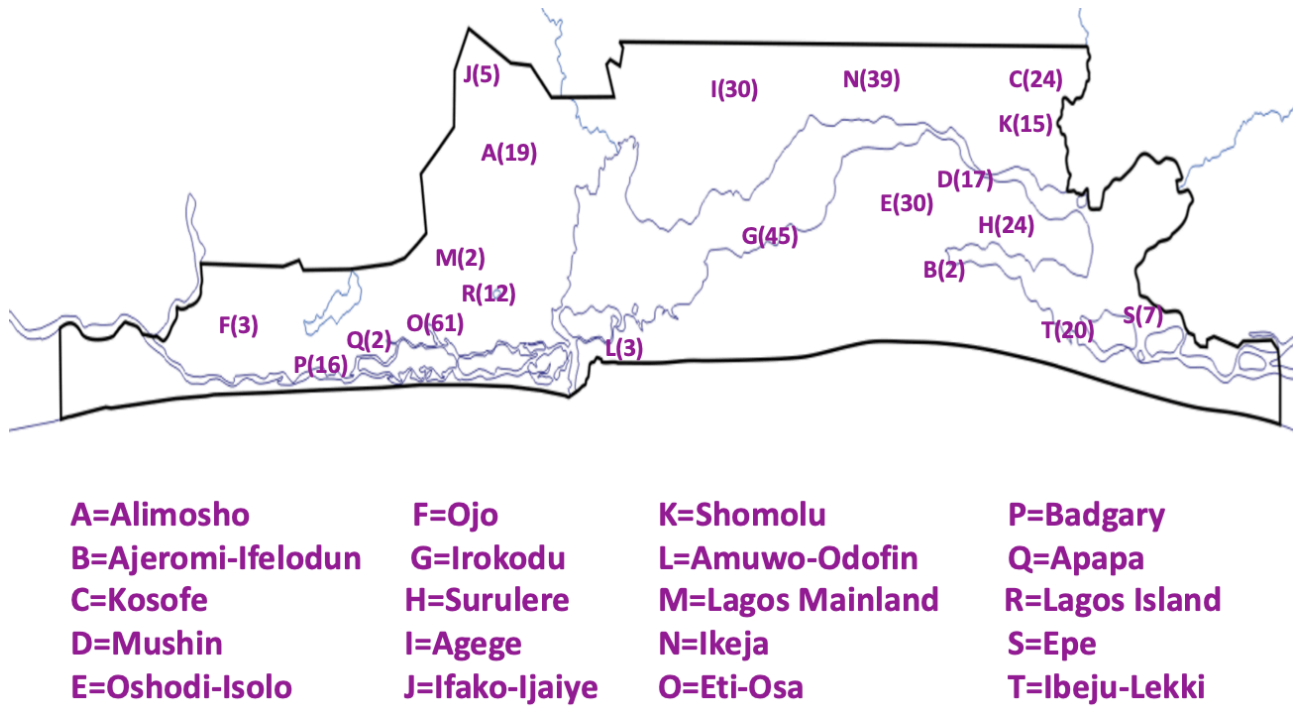


Figure 4. Missing Data Analysis Decision Tree

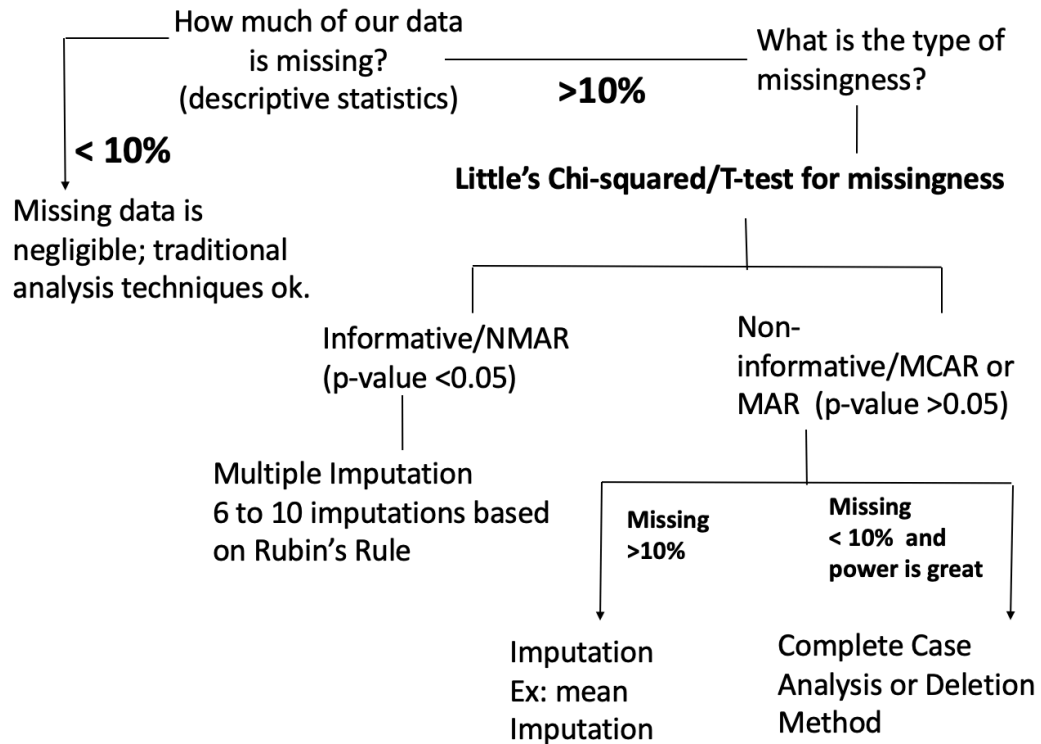


Table 1. Description of variables and percent missingness for LASAMBUS dataset

Variable	n(%)	n(%) Missingness by variable*
Plate Number	---	106(7)
Location call was made LASTMA Hospital/Clinic Other	1 (0.16) 3(0.47) 640(99.38)	887(39)
Age	36.08(12.24)	1054(69)
State of Origin	See Figure 2	1065(70)
Gender Male Female	373(73.28) 136 (26.72)	1022(64)
Occupation information present? Yes No	411(26.85) 1120(73.15)	1120(73)
Local Government Areas (LGA)	See Figure 3	1164(76)
Triage Red Yellow Green Black	4(0.26) 6(0.39) 25(1.63) ----	1496(98)
Temperature (°C) mean(SD)	37.04(6.65)	94
Pulse		63
SPO ₂ mean(SD)	95.88(7.35)	83
Blood pressure mean(SD) Systolic Diastolic	124.96(21.82) 78.39(14.77)	67 67
Respiratory Rate mean(SD)	4.92(0.47)	1148(75)
GCS score Eye None To pain To speech Spontaneously	28(6.75) 10(2.41) 10(2.41) 367(88.43)	1116(73)
GCS Verbal Response None Incomprehensible speech	23(5.56) 12(2.90) 7(1.69)	1117(73)

Inappropriate Confused speech Oriented	17(4.11) 355(85.75)	
GCS Motor Response None Extension Abnormal flexion Withdrawal Localizes pain Obeys command	15(3.62) 6(1.45) 9(2.17) 8(1.93) 17(4.11) 359(86.71)	1117(73)
Total Trauma Score mean (SD)	11.47(1.66)	1153(75)
Outcome Responded, no crash (false call) Responded, crash was already addressed Responded, addressed crash Did not respond Other	373(25.88) 312(21.65) 528(36.64) 21(1.46) 207(14.37)	90(6)

*Percent missingness rounded to nearest whole number

Table 2. Little's Chi-Squared Test for MCAR/MAR

	State of Origin	LGA	Age	Gender	RTA
Occupation information recorded?	0.006	0.001	0.057	0.975	0.000
Systolic BP	0.299	0.036	0.000	0.165	0.000
Diastolic BP	0.443	0.031	0.001	0.269	0.000
Respiratory Rate	0.143	0.128	0.046	0.805	0.000
GCS eye score	0.151	0.126	0.330	0.975	0.000
GCS verbal score	0.151	0.135	0.330	0.927	0.000
GCS motor score	0.175	0.126	0.320	0.927	0.000
Systolic score	0.290	0.141	0.256	0.413	0.000
Total GCS Score	0.306	0.128	0.240	0.507	0.000
Outcome of EMS Call	0.500	0.995	0.978	0.234	0.000

*LGA=Local Government Area; BP=Blood Pressure; GCS=Glasgow Coma Score; RTA=Event reported as a Road Traffic Accident