PLOS ONE

RESEARCH ARTICLE

Mapping risk of ischemic heart disease using machine learning in a Brazilian state

Marcela Bergamini[1], Pedro Henrique Iora[2], Thiago Augusto Hernandes Rocha[3], Yolande Pokam Tchuisseu[3], Amanda de Carvalho Dutra[1], João Felipe Herman Costa Scheidt[2], Oscar Kenji Nihei[4], Maria Dalva de Barros Carvalho[1], Catherine Ann Staton[3,5], João Ricardo Nickenig Vissoci[3,5], Luciano de Andrade[1,2]*

1 Postgraduate Program in Health Sciences, State University of Maringa, Maringa, Brazil, **2** Department of Medicine, State University of Maringa, Maringa, Brazil, **3** Duke Global Health Institute, Duke University, Durham, North Carolina, United States of America, **4** Education, Letters and Health Center, State University of the West of Parana, Foz do Iguaçu, Parana, Brazil, **5** Division of Emergency Medicine, Department of Surgery, Duke University Medical Center, Durham, North Carolina, United States of America

* landrade@uem.br

Abstract

Cardiovascular diseases are the leading cause of deaths globally. Machine learning studies predicting mortality rates for ischemic heart disease (IHD) at the municipal level are very limited. The goal of this paper was to create and validate a Heart Health Care Index (HHCI) to predict risk of IHD based on location and risk factors. Secondary data, geographical information system (GIS) and machine learning were used to validate the HHCI and stratify the IHD municipality risk in the state of Paraná. A positive spatial autocorrelation was found (Moran's I = 0.6472, p-value = 0.001), showing clusters of high IHD mortality. The Support Vector Machine, which had an RMSE of 0.789 and error proportion close to one (0.867), was the best for prediction among eight machine learning algorithms after validation. In the north and northwest regions of the state, HHCI was low and mortality clusters patterns were high. By creating an HHCI through ML, we can predict IHD mortality rate at municipal level, identifying predictive characteristics that impact health conditions of these localities' guided health management decisions for improvements for IHD within the emergency care network in the state of Paraná.

Citation: Bergamini M, Iora PH, Rocha TAH, Tchuisseu YP, Dutra AdC, Scheidt JFHC, et al. (2020) Mapping risk of ischemic heart disease using machine learning in a Brazilian state. PLoS ONE 15(12): e0243558. https://doi.org/10.1371/journal.pone.0243558

Editor: Elena G. Tolkacheva, University of Minnesota, UNITED STATES

Received: April 29, 2020

Accepted: November 23, 2020

Published: December 10, 2020

Data Availability Statement: All relevant data are within the manuscript and can be found following the steps on manuscript methods.

Funding: This study received support in the form of the following grant, awarded to LA and MB: Government Agency - Grant number 028/2017, Programa de Pesquisa para o SUS (PPSUS), 2015 edition, Fundação Araucária-PR / SESA-PR / CNPq / MS-Decit; http://www.fappr.pr.gov.br/sites/fundacao-araucaria/arquivos_restritos/files/

Introduction

Cardiovascular diseases (CVDs) are the leading cause of deaths globally. Among all cardiovascular diseases, ischemic heart disease (IHD) causes the most deaths, both in high-income countries as well as in low- and middle-income countries. More than nine million people died from IHD in 2016 worldwide [1]. Brazil had an average of 306 IHD deaths daily, or one death every five minutes [2]. Cardiac diagnostic tests availability are limited, mostly occurring in urban centers [3], disproportionately limiting the accessibility of remote population and those from small cities.

A vast amount of research on risk factors predicts negative outcomes for patients with IHD. Most studies addressing the prediction of IHD mortality are related to the individual after the

documento/2020-06/2017_atode_031_resultadocp0116.pdf.

Competing interests: The authors have declared that no competing interests exist.

event and use secondary data collected from medical records [4,5]. However, models based on individual data do not represent cultural and socioeconomic disease determinants, which are essential for public health policy and population well-being. The relationship between municipality indicators with socioeconomic and demographic factors, health coverage, and high mortality IHD rates is well documented in the scientific literature [6–8].

Traditional methodological approaches using linear models are usually unable to show subtle associations between indicators and mortality rates or provide insights into factors that can affect health management [9–12]. However, few methodological validated studies seek to predict IHD mortality rates using a data-based innovation that stratifies municipal IHD mortality risk as a public health planning strategy. No machine learning studies predict mortality IHD rate using municipal data.

Using machine learning to generate a risk score allows the use of a greater number of variables at the same time, achieving greater sensitivity and specificity and providing robust results [5,13–15]. Machine learning is better than calculating an index from basic mathematical functions with indicators chosen by experts because it reduces methodological bias risks, enables the discovery of previously unknown regularities and is more reliable in decision-making [8,16–18].

The objective of this study was to create and validate the Heart Health Care Index (HHCI), using a machine-learning algorithm to stratify the IHD municipal risk in the state of Paraná, Brazil. With this index, it will be possible to propose actions and targets to reduce regional disparities. This tool also serves as a source of information for the redesign of actions in IHD care within the emergency care network in the Brazilian state of Paraná.

Materials and methods

Study design and location

This is an observational, cross-sectional, ecological study that used secondary data, spatial analysis and machine learning techniques to create and validate the Heart Health Care Index (HHCI). Once validated, the HHCI was used to stratify the municipal risk for IHD from people who are 20 to 79 years of age, in the 399 municipalities of Paraná state.

Paraná has more than 11.34 million inhabitants (5.44% of the Brazilian population) and is located in the southern region of the country. Municipalities are distributed into 10 administrative regions, with about 90% of them having fewer than 50,000 inhabitants [19]. Paraná ranks among the top five in Human Development Index (HDI: 0.749) among all the 27 Brazilian federative units, being classified as high HDI [20].

This study was approved by the Ethics Committee of State University of Maringá under no. 2.232.580 and was exempted from informed consent.

Source data and study variables

Population, socioeconomic and demographic information as well as the georeferenced cartographic base of Paraná state were obtained from the Brazilian Institute of Geography and Statistics (IBGE) [19]. Information of health coverage was obtained from Ambulatory Information System Database (SIA/DATASUS) [http://www2.datasus.gov.br/DATASUS/index.php?area=0202&id=19122] and IHD mortality was obtained from the Mortality Information System of the Ministry of Health (SIM/DATASUS) [http://www2.datasus.gov.br/DATASUS/index.php?area=0205&id=6937] and expressed in relative and absolute values. In this study, the variables were collected at the municipality level. Our analysis compared rates of IHD by municipalities, seeking to fit a prediction model from municipality level socioeconomic, demographic and health coverage information.

IHD mortality. We selected IHD cases using the International Statistical Classification of Diseases and Related Health Problems–10th Revision (ICD-10), specifically as codes I20 to I25 [1]. To minimize possible mortality-related data fluctuations, the average mortality rates per 100,000 inhabitants (age-adjusted) for the 2009–2014 period were calculated as the outcome for algorithm learning and the 2015 mortality rate was used for model validation [21]. Mortality rates were smoothed through the Empirical Bayesian Estimator for each municipality in Paraná state using GeoDa™ software [22] to reduce extreme behaviors in different size populations. Predictive variables were composed of socioeconomic, demographic and health coverage indicators for each of the 399 cities.

Municipal socioeconomic and demographic indicators. Socioeconomic and demographic: 1. Gross domestic product (GDP); 2. Municipal Human Development Index (MHDI); 3. Expectation of years of child study; 4. Illiteracy rate; 5. Percentage of people with elementary school; 6. GINI index; 7. Average household income per capita, 8. Total employed population; 9. Informal employment rate; 10. Unemployment rate; 11. Income ratio; 12. Percentage of the population with low-income by municipality; 13. Aging rate; 14. Demographic density; 15. Degree of urbanization.

Municipal health indicators. Health management conditions: 1. Municipalities were classified in seven strata according to their size and features: small (unfavorable, regular, favorable), medium (unfavorable, regular, favorable) and large, in which similar municipalities were grouped as unfavorable, regular, favorable, based on 14 indicators with four basic categories (demographic, socioeconomic, health and structural service) [23].

Health coverage indicators: The following variables were used: 1. Family health strategy coverage ratio (%); 2. Rates of consultation for diabetic population; 3. Rates of consultation for hypertensive population; imaging tests per 10,000 inhabitants (4. Cardiac catheterization; 5. Myocardial scintigraphy; 6. Echocardiography); 7. IHD morbidity rate per 100000 inhabitants; 8. Cardiologists; 9. Hemodynamic laboratories, ambulances (10. basic and 11. advanced); 12. Spatial accessibility indexes.

A spatial accessibility index, ranging from 0 (poor index) to 1 (ideal index), measured the proximity and availability of sufficient provision of care for the population. This index was generated using the *two-step floating catchment area (2SFCA) method*. This method has two steps, where catchment areas are created by using 60-minute buffers generated by service area (network analysis) from distance and highways speed using software ArcMap (version 10.5). For the first step, catchment areas from the provider to population were created and extension of the proportion calculated from the centroids is verified in the second step [24,25].

Data analysis

Geospatial analysis. First, choropleth maps were generated with IHD smoothed mortality rates divided into quantiles ranges in the municipalities studied. The exploratory spatial data analysis (ESDA) was used to determine the measurements of Global Spatial Autocorrelation (Moran's I) and Local Indicators of Spatial Association (LISA) using GeoDa software (version 1.12.0, 2017) and QGIS (version 2.14.20).

HHCI development. Machine learning models (Fig 1) were used and constructed according to the TRIPOD guideline [26].

Pre-processing. All the variables were initially included in one of the three strata (socioeconomic, demographic or health coverage) with no greater or less pre-hypothesis importance. The caret package [27] had functions integrated in RStudio (version 3.4.4) and was used for training and prediction functions of essential algorithm. The *doSNOW* package was used for parallel multiprocessing. Thirty-seven variables were registered and pre-processed; two

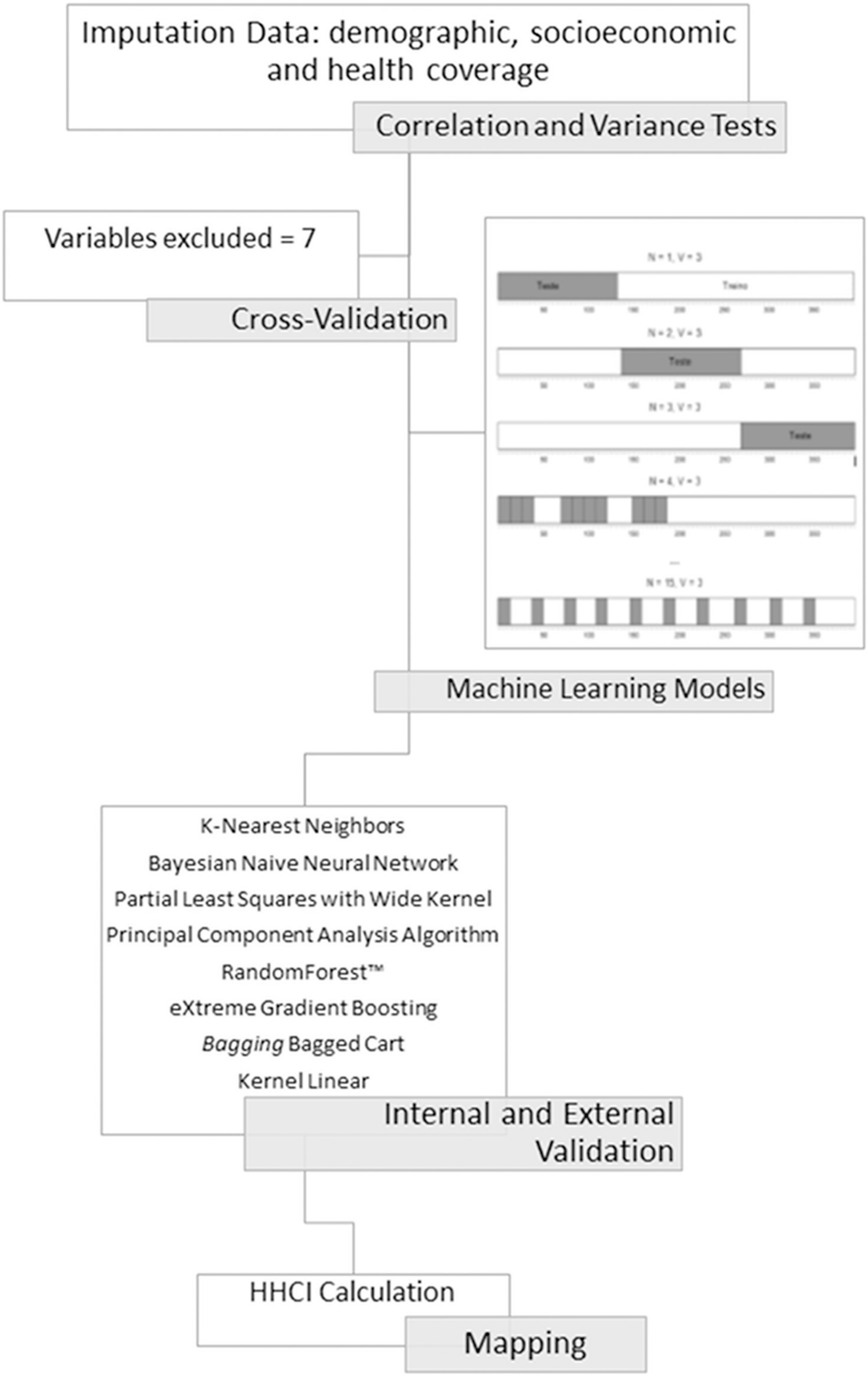
Imputation Data: demographic, socioeconomic and health coverage
Correlation and Variance Tests
Variables excluded = 7
Cross-Validation
N = 1, V = 3
Teste
Treino
N = 2, V = 3
Teste
N = 3, V = 3
Teste
N = 4, V = 3
...
N = 15, V = 3
Machine Learning Models
K-Nearest Neighbors
Bayesian Naive Neural Network
Partial Least Squares with Wide Kernel
Principal Component Analysis Algorithm
RandomForest™
eXtreme Gradient Boosting
Bagging Bagged Cart
Kernel Linear
Internal and External Validation
HHCI Calculation
Mapping

Fig 1. Representative flowchart of the main machine learning model development stages. 1) Variables pre-processing (correlation and variation tests); 2) Predictive variables cross-validation; 3) Machine learning models discrimination and validation (internal and external validation); 4) Heart Health Care Index (HHCI) calculation and mapping.

https://doi.org/10.1371/journal.pone.0243558.g001 identification variables were removed (IBGE codes and city names). The scale native algorithm centralized the Xn numeric variables and these were staggered to X'n. Pearson correlation identified multicollinearity for each of the variables pairs among the 35 remaining indicators (correlation matrix 35x35). A near-zero variance algorithm was applied to the variables, but no significant near-zero variance was found [27].

Cross-validation. Cross-validation was performed using predictive variables as a training parameter for the mean of the period 2009–2014. The data (n = 399) was separated into three equal blocks; two were used for the training algorithm and the rest were used for internal validation. This process was repeated 15 times, having a different validation block each time. The outcomes of this process are used to optimize the algorithm trained for independent database future predictions. For each repetition, one of the three blocks were used for validation. The observed mortality rate was just presented to the model for comparison with the predicted one.

Models discrimination. Eight machine learning models were tested and compared according to their mortality rate predictive accuracy [28]. The final predictive model was chosen from the following algorithms: K-Nearest Neighbors Algorithm ("knn" in CRAN-R *caret* package v.6.0–80) [27]; Bayes Theorem Algorithms: Bayesian Naive Neural Network ("brnn" in CRAN-R *brnn* package v.0.7) [29]; Dimensional Reduction Algorithms: Partial Least Squares with Wide Kernel Algorithm e Principal Component Regression Algorithm ("kernelpls" e "pcr" in CRAN-R *pls* package v.2.7–0) [30]; Decision Tree Models: RandomForest™ ("rf" in CRAN-R *randomForest* package v.4.6–14 May 2018) [31], por *Bagging* Bagged Cart ("TreeBag" in CRAN-R *caret* package v.6.0–80) [27] e eXtreme Gradient Boosting ("xgbDART" in CRAN-R *xgboost* package v.0.71.2) [32]; Vector Support Machine Algorithm with Kernel Linear ("svmLinear" in CRAN-R *kernlab* package v.0.9–27) [33].

Root Mean Square Error (RMSE) between observed mortality rate and predicted mortality rate was used to evaluate the performance of models for each of the 399 municipalities. The RMSE is calculated according to the formula:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (P_{\mathrm{i}} - O_{\mathrm{i}})^2}{\mathrm{n}}}$$

where P = predicted mortality rate, O = observed mortality rate, n = number of municipalities. This is a bank/variable dependent metric.

A lower RMSE indicates smaller mean differences between observed mortality rate and predicted mortality rate.

Models validation. The RMSE has no reference values because it depends on the dimension of results variables where greater accuracy is represented by smaller RMSE values [34]. However, RMSE usually presents low values when overfitting occurs, which indicates the convergence of the observed mortality rate and predicted mortality rate trending towards 100% approximation [35]. The phenomenon of overfitting can occur when training data are exposed to machine learning models. Overfitting is defined as an exceptional fit to the training data but is inaccurate when predicting unknown values. It is detected when the algorithm prediction training has near-perfect accuracy.

To detect models overfitting and adequacy, two exclusive datasets not included in the training phase validation were created. The 2014 and 2015 mortality rates databases were used as independent datasets to test the trained machine learning algorithm external validity.

The 2014 data are intermediate between the average of the years 2009–2014, unknown to the model when isolated. The 2015 dataset presents values never seen before by the algorithm. The indicators values and 2015 rate follow the trend pattern with 2014 and with the previous years, and thus could be used to validate and adapt the over-adjusted models or non-over-adjusted models.

Model calibration. Model calibration was performed comparing the validation on 2015 data to check for RMSE stability and exclude randomly generated results. Additionally, the performed procedure aimed to avoid tuning specific model parameters biases from model to model. The calibration plot was used to assess the predicted versus the actual values regarding each unit of analysis.

Heart health care index calculation. Prediction is the secondary goal of HHCI outcome. Through the predicted mortality rate compared to the observed mortality rate, it is possible to define for each municipality an index of health attention.

Each of the models, with a different mathematical logic, generated a complex object of calculation that presented each of the variables associated with a weight coefficient (WC). Thus, predicted mortality rate (PMR) can be expressed and obtained by the formula:

$$\mathrm{PMR_n} = \mathrm{W_1 C_{1_n}} + \mathrm{W_1 C_{1_n + \ldots} W_{28} C_{28n}}$$

The predicted mortality rate (PMR) of the chosen model is again scaled within the range 0 to 1 and its values are adjusted according to the formula:

$$\mathrm{MR_{adj(0-1)}} = \frac{\mathrm{MR_p} - \min(\mathrm{MR_p})}{\max(\mathrm{MR_p}) - \min(\mathrm{MR_p})}$$

where, min = minimum and max = maximum.

$\mathrm{MR_{adj(0-1)}}$ is the complement of the index and represents spatially adjusted and interval-graded smoothed mortality. This value represents a negative outcome, indicating that higher values of $\mathrm{MR_{adj(0-1)}}$ represents lower access to health for IHD.

Index (I) is obtained according to:

$$\mathrm{I_n} = 1 - \mathrm{MR_{adj(0-1)n}}$$

A choropleth map with spatial distribution of HHCI was generated.

Results

IHD mortality in the state of Paraná

IHD deaths numbered 32310 from 2009–2015 in the state of Paraná for a 20–79 year old population, predominantly male (64.2%), with 1 to 3 years of elementary schooling (31.4%), white (78.9%) and married (54.6%).

The exploratory analysis of spatial IHD mortality rates (Fig 2A) using the Moran Global and LISA Analysis showed a strong spatial dependence, identifying significant spatial clusters (I = 0.6472, p = 0.001) with high IHD mortality rates municipalities surrounded by high rates municipalities (Fig 2B).

Pre-processing

All predictive variables were tested for linear correlation using Pearson's *r* index. The test was performed by generating all possible pairs of 35 variables and calculating the *r* value of each pair. Pairs with an *r* value equal/greater than 0.7 or equal/smaller than -0.7 were found and separated, for a total of 20 [36].

As a result, seven variables were excluded: five socioeconomic (average household income per capita, MHDI, percentage of inhabitants in households with no complete elementary education, occupation rate and income ratio), one demographic (degree of urbanization) and one

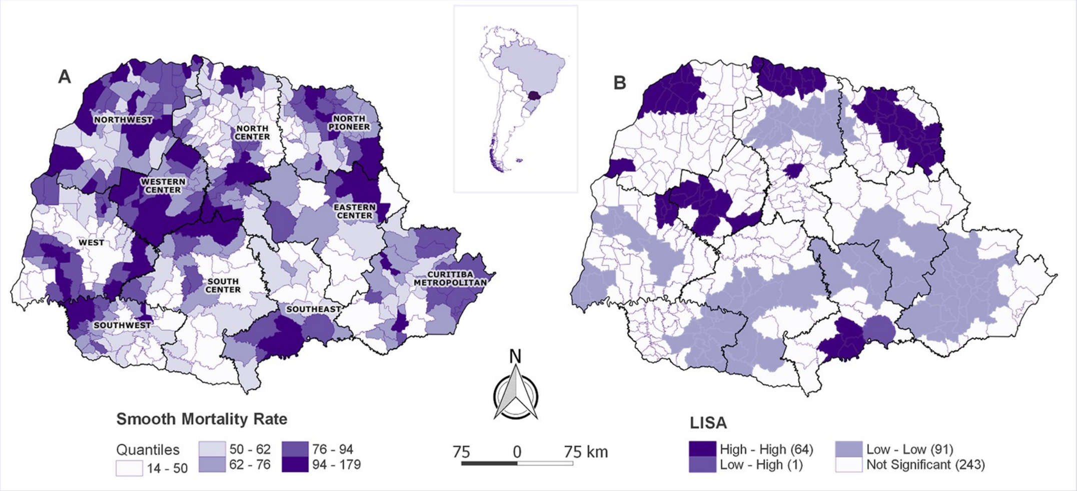

Fig 2. Spatial exploratory analysis of municipalities' IHD mortality rates. A—Distribution of spatially IHD mortality rates observed in the state of Paraná from 2009–2015 in quantiles by municipality. B—Local Indicators of Spatial Association (LISA) of IHD mortality rates by municipality.

https://doi.org/10.1371/journal.pone.0243558.g002

related to access (index of accessibility to cardiologists) for having higher mean absolute Pearson's *r* value in all their composed pairs.

Machine learning models calibration for IHD prediction

The best algorithms presented smaller and more consistent RMSE variability between predicted and observed mortality rates. Data variability was measured using the error proportionality (close to 1) being the larger error the lower proportion index (Table 1).

The calibration was evaluated comparing the predicted and the observed values as indicated in the Fig 3, and a suitable adjustment was confirmed between the predicted and the observed outcome variables. Three machine learning models presented more robust results to predict IHD mortality rates for 100000 inhabitants presenting the lowest RMSE values and lower

Table 1. Calibration of tested models ordered according to the performance indicator (RMSEs and proportionality between the RMSEs).

MODELS	RMSEs			PROPORTIONALITY	
	2009–2014	2014	2015	2014	2015
XGB	0.299	0.835	0.877	0.358	0.341
RF	0.308	0.832	0.894	0.37	0.345
BRNN	0.491	0.945	1.059	0.52	0.464
Tree Bag	0.551	0.871	0.871	0.633	0.633
PLS	**0.771**	**0.912**	**0.932**	**0.845**	**0.827**
PCR	**0.786**	**0.917**	**0.919**	**0.855**	**0.855**
SVM	**0.790**	**0.910**	**0.940**	**0.867**	**0.840**
KNN	0.842	0.915	0.916	0.92	0.919

BRNN–Bayesian Naive Neural Network; KNN–K-Nearest Neighbors Algorithm; PCR–Principal Component Regression Algorithm; PLS–Partial Least Squares with Wide Kernel Algorithm; RF–Random Forest; SVM–Support Vector Machine; TreeBag; XGB–xbgDART/RMSE–Root Mean Square Error of Prediction.

https://doi.org/10.1371/journal.pone.0243558.t001

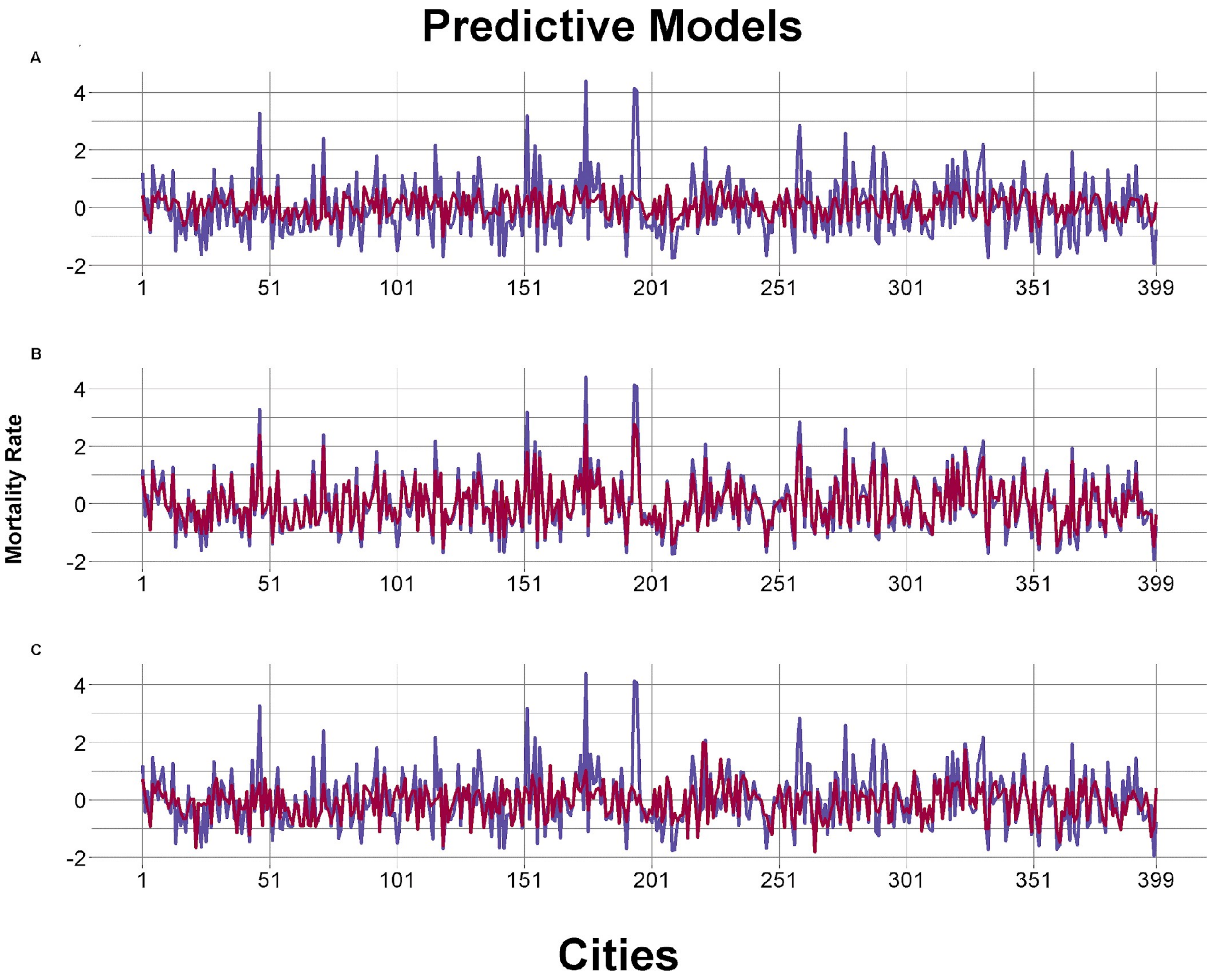

Fig 3. Calibration graphs of the tested predictive models (adjustments using RMSE). A- Example of underfitting calibration model with the worst adjustment (K-Nearest Neighbors); B- Example of overfitting calibration model (Random Forest); C: Example of best fit model (Support Vector Machine). Blue represents the observed mortality rate and red represents the predicted one.

https://doi.org/10.1371/journal.pone.0243558.g003

variability during the process: Partial Least Squares with Wide Kernel Algorithm (PLS), Principal Component Regression Algorithm (PCR) and Support Vector Machine (SVM) (Fig 3C).

Some models were possibly overfitted (Bayesian Naive Neural Network [BRNN], Random Forest [RF], TreeBag and xbgDART [XGB]), performing very well in the first part of data analysis (low RMSE values) but inconsistently in the cross validation process (Fig 3B). The K-Nearest Neighbors Algorithm (KNN model), presenting a good RMSE, underwent the underfitting phenomenon obtaining practically the same prediction for all 399 municipalities with no real variability for each one (Fig 3A).

Sensitivity tests were performed to validate and decide the performance of the models based on the validation criteria and consistency of the prediction errors obtained by the best performed models (SVM and PCR). Simulations were performed to confirm the results and observe the robustness of the outcomes in each model and to evaluate the error variation range.

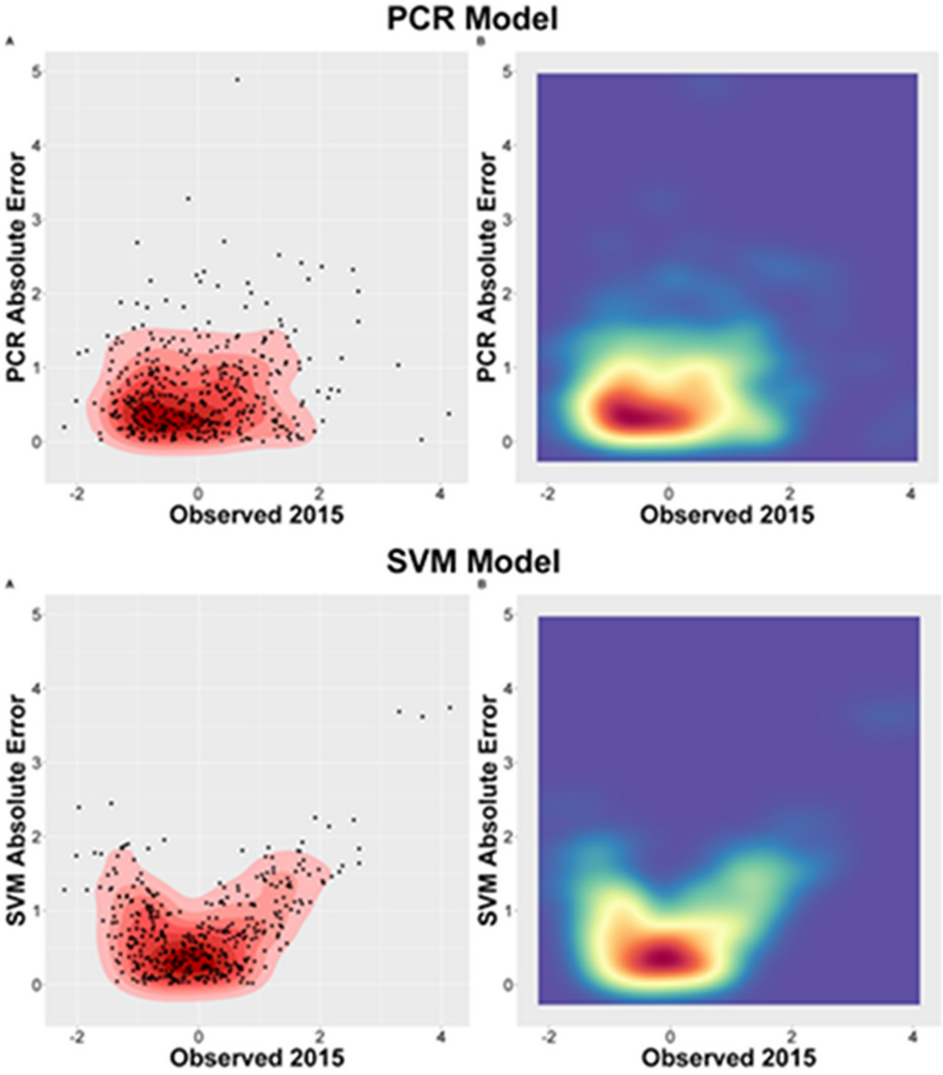

Fig 4. RMSE distribution representation. A—PCR model (2015) B—SVM model (2015).

https://doi.org/10.1371/journal.pone.0243558.g004

The prediction error presented different distribution patterns based on the observed mortality rate value, even though the mean error (RMSE) was approximately identical in both models. The SVM model presented smallest errors concentration when the mortality rate approached its mean value and increasing errors as it approached the rate limits. In contrast, the PCR model presented a less homogeneous distribution pattern of its errors. The variation showed greater independence in relation to the predicted value and did not demonstrate a clear correlation with the observed value, being able to be an outcome of prediction that generalizes predicted values. To demonstrate the difference between the models, a density plot was generated (Fig 4). The different distributions of the error in both models are reflections of the predictive mechanism of each algorithm.

The weight for the SVM model variables are presented on S1 Fig. The top five positive related variables were aging rate, illiteracy, family health strategy coverage ratio, rates of consultation for diabetic population and municipal classification (small [regular]) and the top five

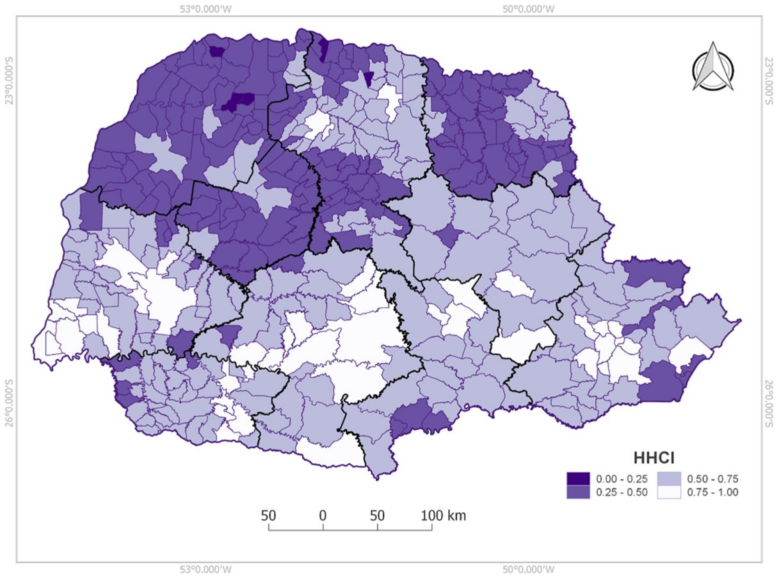

Fig 5. Distribution of the generated municipalities index. 1 lower risk and 0 higher risk.

https://doi.org/10.1371/journal.pone.0243558.g005

negative related variables were advanced ambulances, GINI index, spatial accessibility index: hemodynamic laboratories, myocardial scintigraphy and basic ambulances.

Model validation

The Support Vector Machine model was subjected to 30 repetitions of train-testing and prediction on 2015 data. The SVM model scored an average RMSE of 0.7904 with standard error of 0.0004 and ranging from 0.7901 to 0.7915. The best tune length value was determined to be 10. The C-value, a specific parameter for SVM models, was determined by the model through repetition of random values and selection of the one that produced the smallest RMSE in observed versus predicted values. In the lowest RMSE produced through repetition, C-values equaled 0.3694.

Heart health care index

The SVM model predicted mortality rate was transformed into a standardized score of 0 to 1 where 1 is lower risk and 0 a higher risk. The municipalities were classified according to the risk of IHD mortality.

Spatial distribution of the scores was plotted in a map showing the lowest indexes (greater risks) are in Northwest, Western Center, Pioneer North (large part) and Central North (South part) (Fig 5).

Discussion

This was the first study to predict the IHD mortality rate in a Brazilian state´s municipalities considering socioeconomic, demographic and health accessibility indicators using machine learning to generate a heart health care index.

The variables aging rate, illiteracy, hemodynamic accessibility, IHD morbidity rate and advanced ambulances were the most important variables for HHCI formation showing how these factors may impact municipal IHD mortality.

Among the eight different decisions that machine learning models tested, three were seen with great potential for IHD mortality prediction. The SVM represented the lowest RMSE maintaining proportionality close to one [37] and more homogeneous concentration of errors showing higher correlation between observed and predicted model. A supervised model was chosen, having less learning independence but presenting an easier results interpretation. This model is widely used to make predictions [13] and allows the weights extraction of each variable to generate the index.

Aging rate was expected to be related to IHD mortality. However, illiteracy and accessibility indexes that are often discussed in scientific literature but not specifying its real importance and weight [3,7,8] were important for index formation and can be considered in public policies to reduce mortality. Examples are increasing the number of ambulances as well as the differential management of patients who are more than 120 minutes from the reference hemodynamics centers.

Health indicators weights directs actions to reduce heart health disparities, creating imperative strategies on public health education and awareness besides aging healthily [38–40] that is a big challenge for public managers when small and rural cities are part of the overall picture [41]. Currently, smaller and rural municipalities have less accessibility and longer waiting time after knowing a cardiac event has occurred, time to ambulance arrival and relocation of this individual to a reference hemodynamic center, generating higher risks of morbidity and mortality.

Only seven from ten administrative regions have an interventional cardiology center. North Paranaense is the exception; for all others, high-high IHD mortality clusters were observed where there are no high complexity interventional cardiology centers or at places further from the reference center corroborating previous findings about distance directly related to IHD mortality rates [6,42,43].

HHCI shows that the higher the mortality clusters, the lower the accessibility index. This is confirmed when the inverse is observed in the southern region where lower mortality rates have HHCI close to one indicating better accessibility. Thus, the HHCI and the main identified variables associated with IHD mortality rates may guide the municipalities' health policy managers to improve such specific indicators to reduce population IHD mortality risk, followed by new cycles of evaluation, applying scientific prediction tools to health policy decisions.

Despite the presence of population heterogeneity, since the variables utilized were at municipal level, each municipal population presents specific singularities since each municipality is independently administered and presents specific general socioeconomic and healthcare realities, organization and policies. Thus, our approach is not supposed to be used on the individual level but rather be used as a municipality marker for health care regarding IHD, suitable to support municipal level government policies discussions and restructurations.

Thus, this study can be replicated and tested with socioeconomic, demographic and health coverage variables in different parts of the world, since the data are freely available. Machine learning allows the generation of results reliable to each region according to its characteristics, without incorrect generalizations or assumptions. With this methodology and secondary data, it is possible to have a municipal and state overview that can be monitored and re-evaluated periodically to verify improvements after a proposed plan's execution.

Data availability for the 399 municipalities of the state with free access stands out as a strong point of the study, with no need to impute data.

One of the limitations of the study was the lack of some socioeconomic variable updating; however, the data from the last census provided by Brazilian Institute of Geography and Statistics (IBGE) remains very relevant. Nonetheless, to reduce this limitation, a greater number of variables available annually have been included and accessibility indexes were generated. Another limitation is the impossibility of generalization of the index because the data is only from one particular Brazilian state. Future research with more states or the entire Brazilian territory is necessary for a greater data volume and variability, making possible the generalization of the score.

These well-adjusted models need to be validated in other regions to the long-term objective of reducing the IHD mortality rates and having a municipal evaluation of them. Identifying and recognizing the main municipal determinants impacting health and predicting IHD mortality throughout the country will allow joint and guided decisions in health management to reduce morbidity and mortality rates as well as the generation of a care index which can assess heart health for each city in Brazil.

Conclusion

Machine learning and big data in health area has grown and being accepted. By predicting the IHD mortality risk at the municipal level through the HHCI with freely obtained data, it becomes easier to generalize the model to the country and thus encourage decisions of health managers. This study showed how public data can and should be used for the development of projects that bring interventional results for health improvement.

Supporting information

S1 Fig. SVM model. Variables weight presentation.
(TIF)

Author Contributions

Conceptualization: Amanda de Carvalho Dutra, Catherine Ann Staton, João Ricardo Nickenig Vissoci, Luciano de Andrade.

Data curation: Marcela Bergamini, Pedro Henrique Iora, Thiago Augusto Hernandes Rocha, Amanda de Carvalho Dutra, João Felipe Herman Costa Scheidt, João Ricardo Nickenig Vissoci, Luciano de Andrade.

Formal analysis: Marcela Bergamini, Pedro Henrique Iora, João Felipe Herman Costa Scheidt.

Funding acquisition: Luciano de Andrade.

Investigation: Marcela Bergamini, Thiago Augusto Hernandes Rocha, Oscar Kenji Nihei, João Ricardo Nickenig Vissoci, Luciano de Andrade.

Methodology: Marcela Bergamini, Pedro Henrique Iora, João Felipe Herman Costa Scheidt, Luciano de Andrade.

Project administration: Marcela Bergamini, Oscar Kenji Nihei, Catherine Ann Staton, João Ricardo Nickenig Vissoci, Luciano de Andrade.

Resources: Luciano de Andrade.

Software: Marcela Bergamini, Pedro Henrique Iora.

Supervision: Thiago Augusto Hernandes Rocha, Oscar Kenji Nihei, Maria Dalva de Barros Carvalho, Catherine Ann Staton, João Ricardo Nickenig Vissoci, Luciano de Andrade.

Validation: Pedro Henrique Iora.

Visualization: Marcela Bergamini, Pedro Henrique Iora, Thiago Augusto Hernandes Rocha, Yolande Pokam Tchuisseu, Amanda de Carvalho Dutra, João Felipe Herman Costa Scheidt, Oscar Kenji Nihei, Maria Dalva de Barros Carvalho, Catherine Ann Staton, João Ricardo Nickenig Vissoci, Luciano de Andrade.

Writing – original draft: Marcela Bergamini, Luciano de Andrade.

Writing – review & editing: Yolande Pokam Tchuisseu, Amanda de Carvalho Dutra, Oscar Kenji Nihei, Maria Dalva de Barros Carvalho, Catherine Ann Staton, João Ricardo Nickenig Vissoci.

References

1. WHO, International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10); Version 2016. Disponível online: https://icd.who.int/browse10/2016/en#/I20-I25 (Acesso: 30 de janeiro).
2. DATASUS–Departamento de Informática do SUS. Informações de Saúde, Epidemiológicas e Morbidade: banco de dados. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0203 (Acesso: 14 jul. 2018).
3. Hertz JT, Fu T, Vissoci JR, Augusto T, Rocha H, Carvalho E, et al. The distribution of cardiac diagnostic testing for acute coronary syndrome in the Brazilian healthcare system : A national geospatial evaluation of health access. PLoS One. 2019; 14(1):1–16. https://doi.org/10.1371/journal.pone.0210502 PMID: 30629670
4. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. Eur Heart J. 2017; 38(23):1805–14. https://doi.org/10.1093/eurheartj/ehw302 PMID: 27436868
5. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. SSM—Popul Heal [Internet]. 2018; 4 (November 2017):95–9. Available from: https://doi.org/10.1016/j.ssmph.2017.11.008 PMID: 29349278
6. De Andrade L, Zanini V, Batilana AP, Araujo De Carvalho EC, Pietrobon R, Nihei OK, et al. Regional Disparities in Mortality after Ischemic Heart Disease in a Brazilian State from 2006 to 2010. PLoS One. 2013; 8(3).
7. Buja A, Canavese D, Furlan P, Lago L, Saia M, Baldo V. Are hospital process quality indicators influenced by socio-demographic health determinants. Eur J Public Health. 2015; 25(5):759–65. https://doi.org/10.1093/eurpub/cku253 PMID: 25667156
8. Jahan S, Et A. Technical notes: Calculating the human development indi. Tech notes [Internet]. 2016; 37 (1):14. Available from: http://dev-hdr.pantheonsite.io/sites/default/files/hdr2016_technical_notes_0.pdf.
9. Muller EV, Aranha SRR, Roza WSS da, Gimeno SGA. Distribuição espacial da mortalidade por doenças cardiovasculares no Estado do Paraná, Brasil: 1989–1991 e 2006–2008. Cad Saude Publica [Internet]. 2012; 28(6):1067–77. Available from: http://www.scielo.br/scielo.php?script = sci_arttext&pid = S0102-311X2012000600006&lng = pt&tlng = pt. https://doi.org/10.1590/s0102-311x2012000600006 PMID: 22666811
10. Jordan M.I., & Mitchell T.M. Machine learning: Trends,perspectives, and prospects. N Engl J Med J Med Internet Res PLOS ONE Clin Pharmacol Ther [Internet]. 2015; 360(96):2153–5. Available from: http://science.sciencemag.org.ezproxy.lib.purdue.edu/content/sci/349/6245/255.full.pdf.
11. Awan SE, Sohel F, Sanfilippo FM, Bennamoun M, Dwivedi G. Machine learning in heart failure : ready for prime time. Curr Opin Cardiol. 2018; 33(2):190–5. https://doi.org/10.1097/HCO.0000000000000491 PMID: 29194052
12. Chen Z, Young L, Yu CH, Shiao SPK. A Meta-Prediction of Methylenetetrahydrofolate-Reductase Polymorphisms and Air Pollution Increased the Risk of Ischemic Heart Diseases Worldwide. Int J Environ Res Public Health. 2018; 15(1453). https://doi.org/10.3390/ijerph15071453 PMID: 29996520
13. Alanazi HO, Abdullah AH, Qureshi KN. A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. J Med Syst. 2017; 41(4). https://doi.org/10.1007/s10916-017-0715-6 PMID: 28285459
14. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. J Am Med Informatics Assoc. 2017; 24(1):198–208. https://doi.org/10.1093/jamia/ocw042 PMID: 27189013

15. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty [Internet]. 2018; 33 (8):2358–61. Available from: https://doi.org/10.1016/j.arth.2018.02.067 PMID: 29656964

16. Rothenberg R, Stauber C, Weaver S, Dai D, Prasad A, Kano M. Urban health indicators and indices—Current status. BMC Public Health [Internet]. 2015; 15(1):1–14. https://doi.org/10.1186/s12889-015-1827-x PMID: 25981640

17. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine : are we there yet ? Hear (British Cardiovasc Soc. 2018; 104:1156–64. https://doi.org/10.1136/heartjnl-2017-311198 PMID: 29352006

18. Koprowski R, Foster KR. Machine learning and medicine : book review and commentary. Biomed Eng Online [Internet]. 2018; 17(17):1–10. Available from: https://doi.org/10.1186/s12938-018-0449-9 PMID: 29391026

19. IBGE–Instituto Brasileiro de Geografia e Estatística. Disponível em: https://ww2.ibge.gov.br/home/ (Acesso: abril/2018).

20. PNUD—Programa das Nações Unidas para o Desenvolvimento. Ranking IDHM Municípios 2010. Brasília: PNUD; 2014. Disponível online: http://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html (Acesso: Outubro, 2018).

21. Rouquayrol M. Z. & Filho N. A. Introdução à Epidemiologia—Volume 21—No 4. ed., rev. e ampl. Rio de Janeiro: Guanabara Kongan; MEDSI, 2006.

22. Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. Geogr Anal 2006; 38: 5–22.

23. Calvo MCM, Lacerda JT De, Colussi CF, Schneider IJC, Rocha TAH, Federal U, et al. Estratificação de municípios brasileiros para avaliação de desempenho em saúde. Epidemiol e Serviços Saúde. 2016; 25(4):767–76.

24. Luo W, Wang F. Measures of spatial accessibility to health care in a GIS environment: Synthesis and a case study in the Chicago region. Environ Plan B Plan Des. 2003; 30(6):865–84.

25. Vo A, Plachkinova M, Bhaskar R. Assessing Healthcare Accessibility Algorithms : A Comprehensive Investigation of Two-Step Floating Catchment Methodologies Family. Twenty-first Am Conf Inf Syst [Internet]. 2015;1–12. Available from: aisel.aisnet.org.

26. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015; 13 (1):1–10.

27. Kuhn M., Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, et al. (2018). caret: Classification and Regression Training. R package version 6.0–79. https://CRAN.R-project.org/package=caret.

28. Reddy CK, Aggarwal CC, editors. Healthcare Data Analytics. Boca Raton: Taylor & Francis Group; 2015.

29. Rodriguez, P. P., & Gianola, D. (2016). brnn: Bayesian Regularization for Feed-Forward Neural Networks. R package version 0.6. https://CRAN.R-project.org/package=brnn.

30. Mevik, B., Wehrens, R., & Liland, K. H. (2016). pls: Partial Least Squares and Principal Component Regression. R package version 2.6–0. https://CRAN.R-project.org/package=pls.

31. Liaw A., & Wiener M. (2002). Classification and Regression by randomForest. RNews 2(3), 18–22.

32. Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y., (2018). xgboost: Extreme Gradient Boosting. R package version 0.6.4.1. https://CRAN.R-project.org/package=xgboost.

33. Karatzoglou A., Smola A., Hornik K., Zeileis A., (2004). kernlab—An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1–20. URL http://www.jstatsoft.org/v11/i09/.

34. Roy K, Das RN, Ambure P, Aher RB. Chemometrics and Intelligent Laboratory Systems Be aware of error measures. Further studies on validation of predictive QSAR models ☆. Chemom Intell Lab Syst [Internet]. 2016; 152:18–33. Available from: http://dx.doi.org/10.1016/j.chemolab.2016.01.008.

35. Chen C, Twycross J, Garibaldi JM. A new accuracy measure based on bounded relative error for time series forecasting. PLoS One. 2017;1–23. https://doi.org/10.1371/journal.pone.0174202 PMID: 28339480

36. Beldjazia A, Alatou D. Precipitation variability on the massif Forest of Mahouna (North Eastern-Algeria) from 1986 to 2010. Int J Manag Sci Bus Res ISSN [Internet]. 2016; 5(3):2226–8235. Available from: http://www.ijmsbr.com.

37. Zheng A. Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls. 2015.

38. Mensah GA. Eliminating Disparities in Cardiovascular Health Six Strategic Imperatives and a Framework for Action. 2005 [cited 2018 Nov 24]; Available from: http://www.circulationaha.org.

39. Malta DC, Morais Neto OL de, Silva Junior JB da. Apresentação do plano de ações estratégicas para o enfrentamento das doenças crônicas não transmissíveis no Brasil, 2011 a 2022. Epidemiol e Serviços Saúde. 2011; 20(4):425–38.

40. Olmo MM, Gotsens M, Borrell C, Martinez-beneito MA, Palència L, Pérez G, et al. Trends in Socioeconomic Inequalities in Ischemic Heart Disease Mortality in Small Areas of Nine Spanish Cities from 1996 to 2007 Using Smoothed ANOVA. J Urban Heal Bull New York Acad Med. 2013; 91(1):46–61.

41. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. Int J Environ Res Public Health. 2018; 15. https://doi.org/10.3390/ijerph15122796 PMID: 30544648

42. Vavalle JP, Granger CB. The need for regional integrated care for st-segment elevation myocardial infarction. Circulation. 2011; 124(7):851–6. https://doi.org/10.1161/CIRCULATIONAHA.110.012617 PMID: 21844091

43. Jollis JG, Al-Khalidi HR, Roettig ML, Berger PB, Corbett CC, Doerfler SM, et al. Impact of Regionalization of ST-Segment-Elevation Myocardial Infarction Care on Treatment Times and Outcomes for Emergency Medical Services-Transported Patients Presenting to Hospitals with Percutaneous Coronary Intervention: Mission: Lifeline Accelerator. Circulation. 2018; 137(4):376–87. https://doi.org/10.1161/CIRCULATIONAHA.117.032446 PMID: 29138292