medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Reducing Inequalities Using an Unbiased Machine Learning Approach to Identify Births with the Highest Risk of Preventable Neonatal Deaths.

Antonio P. Ramos *
California Population Center
University of California, Los Angeles, USA
Research Department,
José Luiz Egydio Setúbal Foundation, São Paulo, Brazil

Fabio Caldieraro †
Brazilian School of Public and Business Administration,
Getulio Vargas Foundation, Rio de Janeiro, Brazil

Marcus L. Nascimento
Research Department,
José Luiz Egydio Setúbal Foundation, São Paulo, Brazil

Rafael Saldanha
National Laboratory of Scientific Computing, Petrópolis, Brazil

January 11, 2024

*tomramos@ucla.edu
†fabio.caldieraro@fgv.br

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Background: Despite contemporaneous declines in neonatal mortality, recent studies show the existence of left-behind populations that continue to have higher mortality rates than the national averages. Additionally, many of these deaths are from preventable causes. This reality creates the need for more precise methods to identify high-risk births so that policymakers can more precisely target them. This study fills this gap by developing unbiased machine-learning approaches to more accurately identify births with a high risk of neonatal deaths from preventable causes.

Methods: We link administrative databases from the Brazilian health ministry to obtain birth and death records in the country from 2015 to 2017. The final dataset comprises 8,797,968 births, of which 59,615 newborns died before reaching 28 days alive (neonatal deaths). These neonatal deaths are categorized into preventable deaths (42,290) and non-preventable deaths (17,325). Our analysis identifies the death risk of the former group, as they are amenable to policy interventions. We train six machine-learning algorithms, test their performance on unseen data, and evaluate them using a new policy-oriented metric. To avoid biased policy recommendations, we also investigate how our approach impacts disadvantaged populations.

Results: XGBoost was the best performance algorithm for our task: the 5% births of the highest predicted risk from this model capture more than 85% of the actual deaths. Furthermore, the risk predictions exhibit no statistical differences in the proportion of actual preventable deaths from disadvantaged populations, defined by race, education, marital status, and maternal age. These results are similar for other threshold levels.

Conclusions: We show that, by using publicly available administrative data sets and ML methods, it is possible to identify the births with the highest risk of preventable deaths with a high degree of accuracy. This is useful for policymakers as they can target health interventions to those who need them the most and where they can be effective without producing bias against disadvantaged populations. Overall, our approach can guide policymakers in reducing neonatal mortality rates and their health inequalities. Finally, it can be adapted to be used in other developing countries.

Keywords: Algorithmic bias, Health care, Health inequality, Machine learning, Neonatal mortality, Targeting program.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1 Background

In recent years, many countries have achieved considerable progress in reducing early-life mortality (ELM), and many are in line to achieve the United Nations' Sustainable Development Goals (SDGs). These reductions are important and associated with improved health outcomes[1,2]. However, health disparities remain high, even in countries in line to achieve the SDGs. These disparities may exist among ethnic groups, geographic regions, and levels of education, to mention a few subgroups[3,4]. This is particularly concerning for deaths from preventable causes, where available interventions could be used[5].

International agencies and local policymakers have recognized these disparities among subgroups. The most common approach to identifying high-risk groups has been stratifying mortality rates by subgroups, such as gender, socioeconomic status, and geographic location. While useful for some purposes, these approaches ignore within-group variability, whereas children from the same subgroup may have very different mortality rates. Recent studies showed that within-group variability is higher than between-group variability[6,7].

The decline in mortality rates makes it even more useful to adopt methods that can precisely identify those who still have a high risk of preventable deaths. This is particularly salient when only a fraction of the population can be given the needed intervention because of two factors. First, in many contexts in the developing world, resources are scarce. At the same time, at-risk individuals may demand considerable attention, thus the importance of not squandering resources with those who do not truly need them. Secondly, the smaller the population that can receive an intervention, the more difficult the task of correctly identifying those individuals that should be targeted. In this paper, we develop and explore a new approach. Using a large administrative data set with individual-level information about each birth, we employ machine learning models (ML) to estimate the risk of preventable neonatal death for new unseen births.

In doing so, we aim to aid local health professionals and policymakers identify which children need special attention, not based on preconceived risk factors. We develop a data-driven approach that combines several risk factors and provides digested information to healthcare providers or policymakers about those neonates who need more attention.

This is particularly useful, for example, in Brazilian regions where teams in the public health care system (SUS) may be responsible for 2000 to 3500 individuals[8], as this identification might be very challenging due to the sheer number of patients under their care. The SUS is the world's largest government-run public healthcare system by number of beneficiaries, land area coverage, and affiliated network with more than one million healthcare providers[9]. Based on our methodology, the use of an easy-to-use app [1] could assist healthcare teams on the ground in their targeting strategies by assigning a risk score for each neonate under their care.

Additionally, we apply a new metric to evaluate the performance of the developed machine learning models, which is appropriate to public health professionals and policymakers. Many

[1]We developed an example of such an app and made it available on the Internet at `https://64o4b7-marcus010nascimento.shinyapps.io/tent_app/`

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ML algorithms have their performance judged by criteria such as specificity and accuracy or F1 metrics that are difficult to interpret for policy purposes. Our metric evaluates the usefulness of a given ML algorithm to identify high-risk births from preventable causes.

A critical feature of any life-saving intervention is that it can only save the lives of those who would have otherwise died. Even a "miracle drug" that can counteract any cause of death can only reduce mortality if given to children who, without it, would have died. Because of this, interventions that cannot be given universally must be carefully targeted to those at the highest risk of mortality (absent the intervention) to have an efficient effect. Our method addresses this issue by ranking births by their risk of preventable deaths.

This approach aligns with recent trends in medical and other fields regarding "evidence-based health policies", where medical decisions, including clinical decisions, should be aided by scientific evidence.[10,11,12,13,14]. It is also in line with recent trends in personalized medicine, which is becoming prominent in other fields of medicine and public health, as the risk assignments are estimated at the individual level[15,16,17].

We also address concerns of bias in ML algorithms, given recent literature that shows the potential risk that the application of these methods can be more favorable to privileged populations[18,19,20,21,22]. Our models do not exhibit this behavior and capture similar proportions of preventable neonatal deaths from advantaged and less-advantaged populations.

2 Methods

2.1 Approach

In this research, the unit of analysis is individual birth. We aim to identify the births with the highest neonatal mortality risk from preventable causes. As such, we included all the available information from the administrative databases that contributed to improving the precision of our targeting.

2.2 Data sources

We use administrative databases from the Brazilian health ministry to obtain birth and death records in the entire country from 2015 to 2017 and information about health facilities, professionals, and available equipment. All data is available at `https://datasus.saude.gov.br`. Still, it is organized into three different health information systems: SINASC (Sistema de Informações sobre Nascidos Vivos), SIM (Sistema de Informação sobre Mortalidade), CNES (Cadastro Nacional de Estabelecimentos de Saúde), which we describe below.

SINASC includes all live births in the Brazilian territory, recording epidemiological and administrative information about the mothers and children. SIM, in turn, includes all deaths in the territory, containing epidemiological and administrative information and their circumstances. Fetal deaths are not considered as they are beyond the scope of this paper. Finally, CNES records a snapshot of Brazilian health facilities at a point in time. These systems contain three tables with all live births, deaths, and health facilities information.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To merge SIM and SINASC data, we used the field NUMERODN. It contains a unique number identifying each live birth. Records on SIM contain this information in cases of deaths within the first year since birth. Subsequently, we merged the information with CNES data by the CNES number, a unique identifier for health facilities in both the SINASC and CNES databases. The resulting raw dataset totals 8,829,944 records.

When merging the three databases, we identified and removed duplicated observations in the SIM and SINASC tables to avoid inconsistencies. With the deduplicated tables, deterministic linkages were executed.

In the raw dataset, a few additional treatments were performed. SIM records that were not linked to a SINASC record were not considered. Moreover, we did not consider a few residual records with no birthdate and records in which the difference between the birthdate and the date of death was negative. The resulting cleaned dataset comprises 8,797,968 births and 59,615 neonatal deaths.

2.3 Feature Engineering

Our set of features[23] consists of the following variables: place of delivery, health facility type, maternal age at birth, sex, 1-min Apgar score, 5-min Apgar score, birth weight, gestational age, week of gestation, pregnancy type, delivery type, maternal education, presence of congenital anomaly, maternal ethnicity, antenatal visits, month of first antenatal visit, presentation type, induced labor, professional that assisted the labor, number of previous live births, number of previous fetal losses and abortions, number of previous pregnancies, number of previous vaginal deliveries, number of previous cesarean deliveries. In addition, we have also used marital status and state of birth (the definition and type of each of these features are in Table 1)

We analyze a nominal categorical target variable with three possible outcomes: alive, preventable death[24], and non-preventable death. Among the non-preventable deaths, we have external causes of death and ill-defined deaths. The number of preventable deaths is 42,290, whereas the number of non-preventable deaths is 17,325.

To improve analysis efficiency, categorical variables were stored with codes. We did so by performing a relabeling procedure guided by the data dictionaries issued by the DataSUS, using the package *microdatasus*[25]. We treated missing data via imputation and applied the package *Amelia*[26]. Both packages are available in the R Statistical Software repository[27].

As a pre-processing procedure, we centered and scaled the data, by subtracting the mean and dividing by the standard deviation. We also identified and excluded features with zero or near zero variance. Finally, we filtered out highly correlated features. Details are available upon request.

2.4 Modeling

The final dataset is partitioned into training and test sets: 7,038,375 observations (80.00% of the total) are used to train six different machine learning algorithms, while 1,759,593

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

observations (20.00% of the total) are used to evaluate the performance of our targeting criterion on new unseen data.

We estimated neonatal preventable infant mortality risk for each birth in the data set through flexible ML methods that use the above features. These methods were logistic regression, least absolute shrinkage and selection operator regression (LASSO), elastic-net regularized logistic regression (elastic net), random forest (RF), extreme gradient boosting over trees (XGBoost), and neural networks (NNs). We used the package *caret* available in the R Statistical Software[27] to run the machine learning algorithms.

Logistic regression[28] is the standard estimation of a linear model that estimates the parameters β_j for each feature j to maximize a logistic likelihood function by minimizing the negative log-likelihood. LASSO[29] is essentially an implementation of linear regression that uses a L_2 ($\sum_j \beta_j^2$) norm penalty to regularize or "shrink" the model, preventing overfitting. It is similar to the logistic regression but includes a penalty term equal to $\lambda(2\sum_j \beta_j^2)$, where the parameter λ is a non-negative real number that determines the strength of the regulation.

Elastic net[30] combines L_1 norm ($\sum_j |\beta_j|$) and L_2 ($\sum_j \beta_j^2$) norm penalties to regularize the model. It minimizes the negative log-likelihood plus a penalty term equals to $\lambda(\frac{\alpha}{2}\sum_j |\beta_j| + \frac{1-\alpha}{2}\sum_j \beta_j^2)$, where the parameters α and λ are defined on the unit interval and on the non-negative real numbers respectively. As particular cases, elastic net comprises LASSO regression ($\alpha = 1$) and logistic regression ($\lambda = 0$).

Our application first tested a cross-validation procedure to choose the parameters α and λ in elastic net and the parameter λ in LASSO. However, their performances were not close to the logistic regression. For that purpose, we fixed $\alpha = 0.5$ and $\lambda = 0.001$ in elastic net, and $\lambda = 0.001$ in LASSO. The method *glmnet* was used for all three algorithms.

The methods RF[31] and XGBoost[32] are tree-based algorithms. The simplest tree-based algorithms are classification and regression trees (CART[33]). Both single-tree models recursively group the outcome observations with similar values using cutoff values of the features. Although single-tree models are easy to interpret, their performance is frequently poor and very sensitive to small changes in the input data. By combining several trees, RF and XGBoost methods improve single-tree algorithm performance. The former averages the estimates of a set of trees, each obtained from a random subset of features and trained on a random subset of the observations. The latter also combines several trees, but it initiates with one tree, and new trees are iteratively trained on the errors of the prior set of trees.

As applied by us, in RF, the *ranger* method was employed, and (i) each forest encompasses 500 trees, (ii) the number of variables randomly sampled for each tree split (`mtry`) was set to 5 (the square root of the number of features), (iii) the minimal node size (`min.node.size`) was set to 1, and (iv) we choose the gini index as splitting rule (`splitrule`). In XGBoost, the *xgbTree* method was employed, and (i) the number of iterations for the boosting procedure (`nrounds`) was set to 250, (ii) the learning rate ($\eta \in (0, 1)$ was set to 0.3 to prevent overfitting, (iii) the maximum depth of the trees (`max_depth`) was set to 4, (iv) the proportion of the variables to be considered for tree construction (`colsample_bytree`) was set to the interval (0.6, 1), and the proportion of observations from the training set used for modeling (`subsample`) was set to the interval (0.5, 1).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The NN methods[34] is constituted by an output layer and node layers, including an input layer and one or more hidden layers. The input layer takes the features, and no processing is done. All kinds of processing are executed on the hidden layers and transferred to the output layer. The output layer, in turn, is the final layer, bringing the final value resultant from the learning process in the hidden layers. The nodes, also known as artificial neurons, are connected, and these associations are characterized by their weights, thresholds, and activation functions. Nodes are activated, and data are sent to the next network layer when their outputs exceed a specified threshold value. Otherwise, no data is transmitted to the next layer.

Although we tested specifications with more than one hidden layer using the *mlpML* method, they performed similarly to the neural network with only one hidden layer. Thus, our application employed the *mlp* method, specifying a layer with 25 nodes.

2.5 Performance Metrics

For our task, we did not find it useful to adopt traditional prediction performance metrics, such as classification accuracy, confusion matrices, specificity/sensitivity statistics, or precision/recall statistics, all of which require a threshold for deciding when a risk score is high enough to merit a warning. These can be misleading when applied to rare outcomes, as in the problem we focus on. In our case, if we predicted no neonatal mortality, that model would be right 99% of the time, yet it would be useless, as it wouldn't allow us to identify those who could be targeted. We neither find it useful to adopt "threshold-free" approaches that report accuracy in a way that does not depend on choosing one threshold, such as ROC-AUC and F-scores do, because they are difficult to give any valuable policy meaning in our context.

We instead recognize that if one has a resource constraint- only a certain fraction of cases one can act on- it gives a reason to compute the proportion of deaths captured by setting the threshold levels of the highest predicted mortality risk. For example, suppose we imagine that a policymaker can only provide intervention to the 5% (or 10%) who need it the most. In this case, the threshold can be set to whatever fraction of high-risk births they have resources for targeting. An appropriate approach, therefore, can concentrate a substantial amount of neonatal deaths in small percentages of high-risk individuals.

2.6 Algorithmic Bias

Algorithmic bias[18,19,20,21,22] is a well-documented problem with striking implications for health care and public policy. Therefore, besides concentrating a substantial amount of neonatal preventable deaths in small percentages of high-risk individuals, our targeting criterion should also be able not to disadvantage the most vulnerable groups.

To check whether our preferred model would not disadvantage the most vulnerable populations, we checked its performance for four different sub-groups identified using the demographic variables in our dataset. These sub-groups are newborns from *non-white mothers*,*low-education mothers*, *underage mothers*, and *single mothers*. We use the test

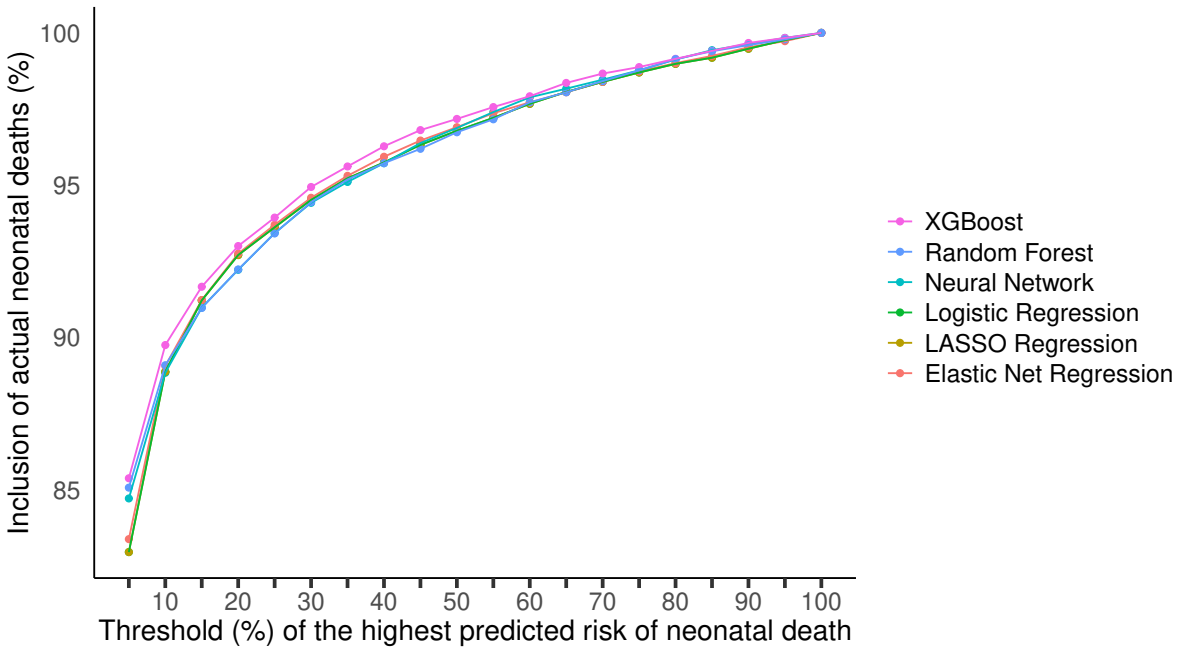
medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

sample as a reference and compare its composition with individuals with the highest predicted risk of neonatal preventable death for different threshold levels. For that, we construct confidence intervals based on the Normal approximation for the mortality rate of each group and check whether these intervals contain their respective mortality rates in the test sample. We also perform hypothesis tests to verify whether the proportions of preventable deaths captured by the algorithm ($\hat{p}$) are statistically equal to the proportion of preventable deaths in the test sample (p_0). The null hypothesis is $H_0 : \hat{p} = p_0$ and the alternative hypothesis, $H_A : \hat{p} \neq p_0$.

3 Results

Recall that we calculate our performance metric by setting the highest predicted mortality risk threshold levels and considering the percentage of neonatal preventable deaths in the test sample for each threshold level. Figure 1 summarizes the results.

Figure 1: Performance of the different ML methods.



Our best model in terms of predictive performance is the XGBoost method. With that algorithm, in our test sample, including the 5% highest risk births, our model captures 85% of preventable neonatal deaths. The XGBoost is never worse than other competing methods and thus is selected as our preferred model.

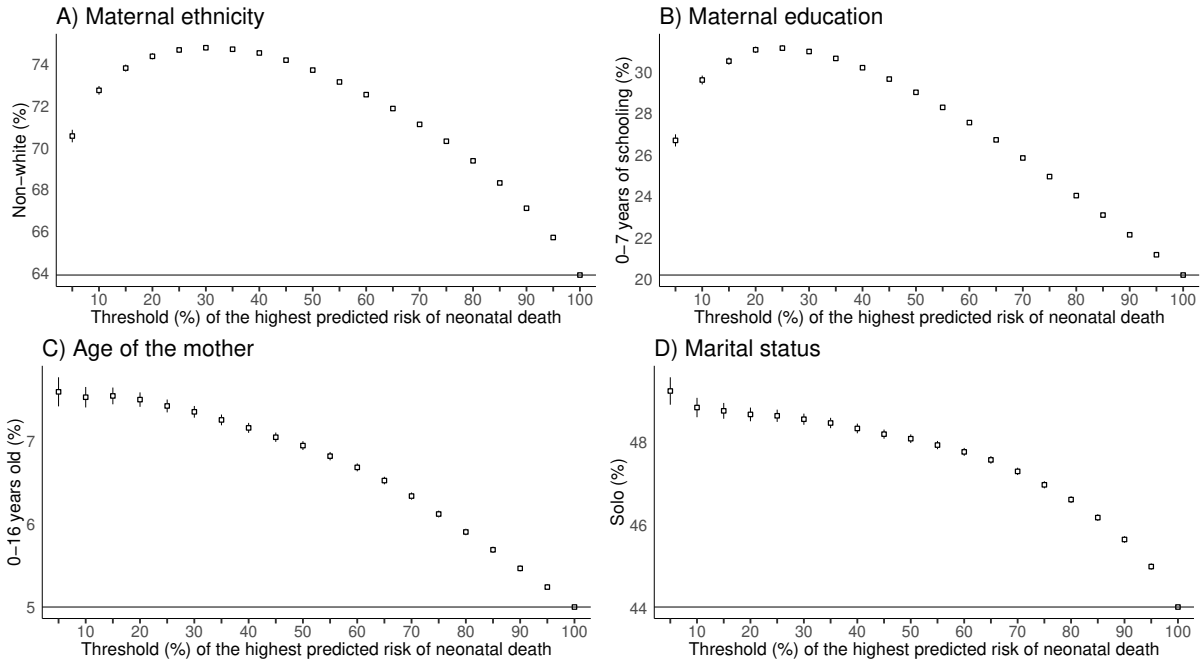
To check whether our preferred model would not disadvantage the most vulnerable populations, we checked its performance for the four different sub-groups of newborns from disadvantaged populations as presented in the Algorithmic Bias subsection.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In the first analysis, we compared the percentage of disadvantaged individuals selected as high-risk versus the proportion of disadvantaged individuals in our test sample for each sub-group. Figure 2 reports these results.

The figure demonstrates that our algorithm selects a significantly higher proportion of individuals from the disadvantaged sub-groups to be high risk for nearly all threshold percentages of the highest predicted risk. Only at the highest percentage thresholds, when the algorithm selects nearly the entire test set, does the proportion of disadvantaged sub-groups converge to the actual proportion in the test set. This means that the proportion of disadvantaged individuals selected by the algorithm is higher than the overall proportion in the test set.

Figure 2: Proportion of individuals selected by the model.

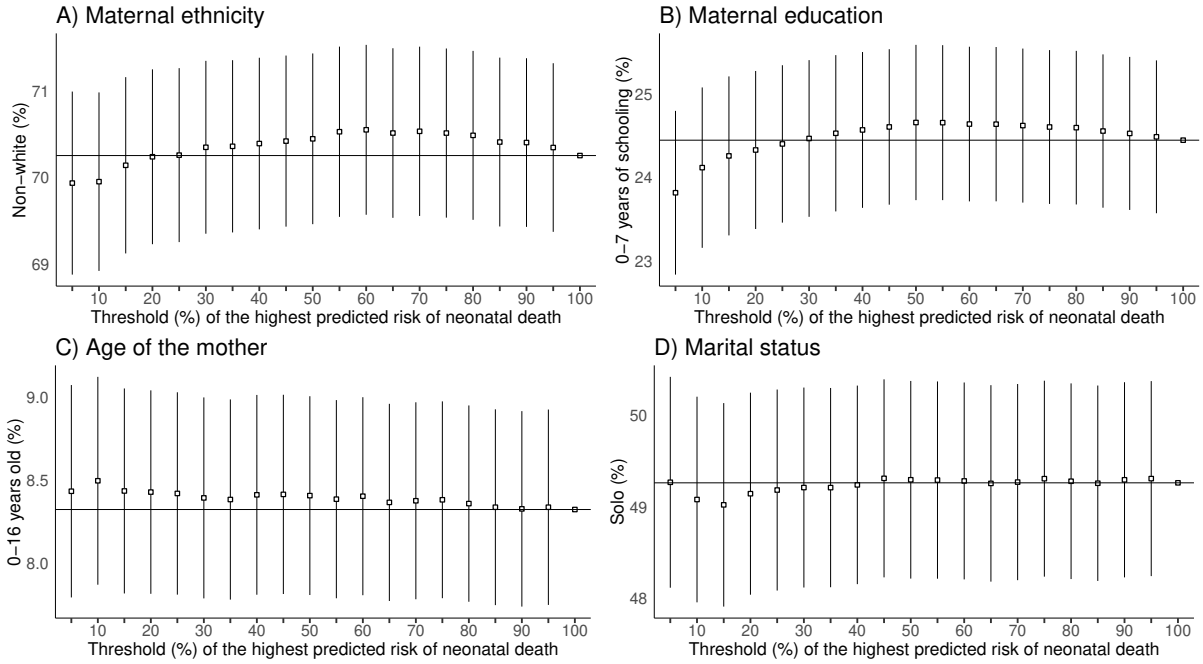


Note: In each graph, the horizontal lines depict the proportion of each ELM in the test set. The data points mark the proportion of actual ELM captured by our model and their confidence intervals.

One wonders whether selecting these individuals would reflect a distortion in the number of preventable deaths captured by the algorithm. Figure 3 depicts the analysis of preventable deaths identified per subgroup. The analysis demonstrates that there are no statistical differences in the proportion of actual preventable deaths from disadvantaged populations that would be included in the selected at-risk births. Therefore, our preferred model is not biased against or favoring underserved groups. On the contrary, using our algorithm would provide a fair inclusion for each population in terms of actual preventable deaths.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 3: Proportion of actual ELM captured by the model.



Note: In each graph, the horizontal lines depict the proportion of each ELM in the test set. The data points mark the proportion of actual ELM captured by our model and their confidence intervals.

To better highlight these results, Table 3 presents the outcomes of the hypothesis tests at a 95% confidence level. The null hypothesis is not rejected for all thresholds and sub-groups depicted in the table. It corroborates that using our algorithm would provide a fair inclusion for each population in terms of actual preventable deaths.

Taken together, these results show that our algorithm selects a higher proportion of individuals from the less-advantaged populations as high-risk births to provide a statistically equal proportion of births with preventable deaths for both disadvantaged and privileged sub-groups.

4 Discussion

Our research puts forward a new analytic approach that integrates large administrative datasets and pairs them with ML models to enable the identification, with a high degree of accuracy, of births with the highest risk of preventable deaths. Furthermore, the approach properly selects a statistically equal proportion of births with preventable deaths for disadvantaged and privileged populations, muffling concerns of algorithmic bias.

We can conceive our approach as a calculator that produces a risk score for each birth. These scores can aid teams of medical professionals to more precisely direct resources to

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

those who have the highest risk of preventable death. As we discussed in the Introduction, doctors and healthcare professionals are often overwhelmed by the number of patients they have to take care of, and this system can be useful to aid them when making decisions about allocating scarce resources.

Our approach is not a replacement for health care professionals, who have subject matter expertise that should not be ignored. Instead, we are offering one additional tool for them. This tool can be particularly useful in a situation such as that in the Brazilian public health systems SUS, in which healthcare teams are responsible for a large number of patients.

Our recommendations are related to a large body of literature in medicine and public health that develops risk scores for individuals to identify those at risk of some event. These scores have been applied to a variety of outcomes[35,36], and our results suggest the possible usefulness of such scores for the identification of neonates with a high risk of preventable deaths. In addition, we also address a common concern: algorithmic biases[18,19,20,21,22]. Our approach is not biased in favor of more privileged groups. Instead, it classifies a higher proportion of individuals from less privileged populations as high-risk births to reach a statistically equal proportion of preventable deaths from more and less privileged groups, alleviating concerns of negative bias in ML algorithms.

It is common to use poverty status as a proxy for at-risk births[37]. Our findings should not be interpreted as recommending against this practice. On the contrary, we implicitly include poverty in our analysis by considering several risk factors that correlate with it.

Our study presumes that healthcare providers can make the right intervention to save at-risk newborns. That is the reason behind ranking the risk of "preventable deaths." Of course, for this to become true, it depends on the actual capacity of the policymaker or healthcare provider to intervene in preventable death cases correctly.

5 Conclusion

Our approach can guide Brazilian policymakers in reducing neonatal mortality rates and health to intervene in preventable death cases correctly and unbiasedly. The methods and metrics developed in this paper have broader applicability and are flexible enough to apply to several scenarios in other developing countries. For example, some countries with incomplete vital registration systems could use surveys like the Health and Demographic Surveys (DHS) instead of administrative data. The risk factors inclusion may also vary across countries, given data availability and political and public health considerations.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data is publicly available. The code used in this analysis will be posted in a public repository.

Competing interests

The authors declare that they have no competing interests.

Funding

The Getúlio Vargas Foundation partially supported this work, award number PAR-004.037.019.00009.

Authors' contributions

APR and FC designed the study and the statistical analysis and wrote the manuscript. MLN performed the statistical analysis and wrote the manuscript. RS assembled the dataset. All authors approved the final manuscript.

Acknowledgements

The authors thank Chad Hazlett, PhD, for his comments on an early version of this paper, and Isabella Grion for her help in editing the paper.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A List of included features

Table 1: Features information: administrative database, type, and description.

Feature	Database	Type	Description
LOCNASC	SINASC	Nominal	Place of delivery (hospital; other health facility; residence; others)
VINC_SUS	CNES	Nominal	Health facility associated with the public healthcare system (yes; no)
IDADEMAE	SINASC	Numerical	Maternal age at birth (in years)
SEXO	SINASC	Nominal	Gender (female; male)
APGAR1	SINASC	Ordinal	1-min Apgar score
APGAR5	SINASC	Ordinal	5-min Apgar score
PESO	SINASC	Numerical	Birth weight (grams)
SEMAGESTAC	SINASC	Numerical	Gestational age (in weeks)
GRAVIDEZ	SINASC	Nominal	Pregnancy type (single; double; triple or more)
PARTO	SINASC	Nominal	Delivery type (vaginal; cesarean)
ESCMAE	SINASC	Ordinal	Maternal education (None; 1-3 years; 4-7 years; 8-11 years; 12 or more years)
IDANOMAL	SINASC	Nominal	Presence of congenital anomaly (yes; no)
RACACORMAE	SINASC	Nominal	Maternal ethnicity (white; black; asian; pardo or mixed; indigenous)
CONSPRENAT	SINASC	Numerical	Number of antenatal visits
MESPRENAT	SINASC	Numerical	Month of first antenatal visit
TPAPRESENT	SINASC	Nominal	Presentation type (cephalic; breech; transversal; other)
STTRABPART	SINASC	Nominal	Induced labor (yes; no)
TPNASCASSI	SINASC	Nominal	Professional that assisted the labor (doctor; nurse; midwife; others)
QTDFILVIVO	SINASC	Numerical	Number of previous live births
QTDFILMORT	SINASC	Numerical	Number of previous fetal losses and abortions
QTDGESTANT	SINASC	Numerical	Number of previous pregnancies
QTDPARTNOR	SINASC	Numerical	Number of previous vaginal deliveries
QTDPARTCES	SINASC	Numerical	Number of previous cesarean deliveries
ESTCIVMAE	SINASC	Nominal	Marital status (single; married; widowed; divorced; common-law marriage)
CODMUNRES	SINASC	Nominal	State of residence

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

B Feature importance

Figure 4: XGBoost feature importance.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

C Predictive performance

Table 2: Predictive performance for preventable neonatal mortality on the test set for each machine learning algorithm.

Performance measure	Machine learning algorithm					
	Neural Network	Extreme Gradient Boosting	LASSO Regression	Elastic Net Regression	Logistic Regression	Random Forest
Balanced Accuracy	0.8450	0.8353	0.8268	0.8302	0.8268	0.8361
AUC	0.9498	0.9532	0.9491	0.9500	0.9491	0.9500
Sensitivity	0.7058	0.6840	0.6658	0.6729	0.6658	0.6850
Specificity	0.9842	0.9866	0.9879	0.9875	0.9879	0.9871
PPV	0.1780	0.1988	0.2100	0.2068	0.2100	0.2051
NPV	0.9985	0.9984	0.9983	0.9984	0.9983	0.9984
Recall	0.7058	0.6840	0.6658	0.6729	0.6658	0.6850
F1-Score	0.2843	0.3081	0.3193	0.3164	0.3193	0.3157

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

D Hypothesis tests

Table 3: Tests for proportions at a confidence level of 95%: comparing the proportion of preventable deaths captured by the algorithm and in the test sample.

Threshold (%)	n	Maternal ethnicity ($p_0 = 70.25\%$)		Maternal education ($p_0 = 24.45\%$)		Age of the mother ($p_0 = 8.32\%$)		Marital status ($p_0 = 49.26\%$)	
		$\hat{p}$ (%)	p-value	$\hat{p}$ (%)	p-value	$\hat{p}$ (%)	p-value	$\hat{p}$ (%)	p-value
5	7,221	69.93	0.5629	23.81	0.2173	8.43	0.7505	49.27	1.0000
10	7,591	69.95	0.5737	24.12	0.5126	8.49	0.7505	49.08	0.7591
15	7,753	70.14	0.8382	24.26	0.7089	8.43	0.7366	49.02	0.6798
20	7,866	70.23	0.9881	24.33	0.8183	8.42	0.7509	49.14	0.8419
25	7,945	70.25	1.0000	24.40	0.9361	8.42	0.7699	49.18	0.8971
30	8,030	70.34	0.8608	24.47	0.9762	8.39	0.8359	49.21	0.9352
35	8,087	70.35	0.8430	24.53	0.8723	8.38	0.8600	49.21	0.9340
40	8,143	70.39	0.7935	24.57	0.8061	8.41	0.7875	49.24	0.9768
45	8,188	70.42	0.7499	24.60	0.7475	8.41	0.7802	49.31	0.9379
50	8,219	70.44	0.7101	24.66	0.6638	8.40	0.7985	49.30	0.9604
55	8,252	70.52	0.5925	24.66	0.6657	8.38	0.8531	49.29	0.9650
60	8,282	70.55	0.5616	24.64	0.6912	8.40	0.8067	49.28	0.9787
65	8,319	70.51	0.6119	24.64	0.6928	8.36	0.9030	49.26	0.9996
70	8,345	70.53	0.5836	24.62	0.7189	8.37	0.8770	49.27	0.9970
75	8,363	70.51	0.6114	24.60	0.7460	8.38	0.8614	49.31	0.9423
80	8,386	70.48	0.6485	24.60	0.7584	8.35	0.9214	49.28	0.9830
85	8,408	70.40	0.7632	24.55	0.8247	8.33	0.9791	49.26	1.0000
90	8,430	70.40	0.7718	24.53	0.8722	8.32	1.0000	49.30	0.9601
95	8,444	70.34	0.8613	24.49	0.9410	8.33	0.9790	49.31	0.9410

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1 Johnson RC, Schoeni RF. The Influence of Early-Life Events on Human Capital, Health Status, and Labor Market Outcomes Over the Life Course. The BE Journal of Economic Analysis & Policy. 2011;11(3).

2 Currie J, Vogl T. Early-Life Health and Adult Circumstance in Developing Countries. Annual Review of Economics. 2013;5:1–36.

3 Smith LK, Manktelow BN, Draper ES, Springett A, Field DJ. Nature of socioeconomic inequalities in neonatal mortality: population based study. BMJ. 2010;341:c6654.

4 Dyer L, Theall KP, Wallace M. Structural racism, racial inequities and urban–rural differences in infant mortality in the US. Journal of Epidemiology & Community Health. 2020;75(8):788–793.

5 Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK, et al. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? Lancet. 2014;384(9940):347–370.

6 Ramos AP, Weiss RE. Measuring Within and Between Group Inequality in Early-Life Mortality Over Time: A Bayesian Approach with Application to India. arXiv preprint arXiv:180408570. 2019;.

7 Ramos AP, Flores MJ, Weiss RE. Leave no child behind: Using data from 1.7 million children from 67 developing countries to measure inequality within and between groups of births and to identify left behind populations. PLOS ONE. 2020;15(10):e0238847.

8 Ministério da Saúde. Portaria nº 2.436; 2017-09-21. Diário Oficial da União.

9 Castro MC, Massuda A, Almeida G, Menezes-Filho NA, Andrade MV, de Souza Noronha KVM, et al. Brazil's unified health system: the first 30 years and prospects for the future. The Lancet. 2019;395(10195):P345–356.

10 Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71–72.

11 Sackett DL. Evidence-based medicine. Seminars in perinatology. 1997;21(1):3–5.

12 Giacomini M. Theory-Based Medicine and the Role of Evidence: Why the Emperor Needs New Clothes, Again. Perspectives in Biology and Medicine. 2009;52(2):234–251.

13 Bluhm R, Borgerson K. Evidence-Based Medicine. In: Gifford F, editor. Philosophy of Medicine. Elsevier; 2011. p. 203–238.

14 Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. The Lancet. 2017;390(10092):415–423.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

15 Hamburg MA, Collins FS. The Path to Personalized Medicine. New England Journal of Medicine. 2010;363:301–304.

16 Hayes DF, Markus HS, Leslie RD, Topol EJ. Personalized medicine: risk prediction, targeted therapies and mobile health technology. BMC Medicine. 2014;125(37).

17 Hoeyer K. Data as promise: Reconfiguring Danish public health through personalized medicine. Social Studies of Science. 2019;49(4):531–555.

18 Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. Journal of Global Health. 2019;9(2):020318.

19 Obermeyer Z, Vogeli BPC, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;60(6464):447–453.

20 Kakani P, Chandra A, Mullainathan S, Obermeyer Z. Allocation of COVID-19 Relief Funding to Disproportionately Black Counties. JAMA. 2020;324(10):1000–1003.

21 Akter S, McCarthy G, Sajib S, Michael K, Dwivedi YK, D'Ambra J, et al. Algorithmic bias in data-driven innovation in the age of AI. International Journal of Information Management. 2021;60:102387.

22 Mullainathan S, Obermeyer Z. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. Quarterly Journal of Economics. 2022;137(2):679–727.

23 Batista AFM, Diniz CSG, Bonilha EA, Kawachi I, Filho ADPC. Neonatal mortality prediction with routinely collected data: a machine learning approach. BMC Pediatrics. 2021;21(322).

24 Malta DC, Duarte EC, de Almeida MF, de Salles Dias MA, de Morais Neto OL, de Moura L, et al. Lista de causas de mortes evitáveis por intervenções do Sistema Único de Saúde do Brasil. Epidemiologia e Serviços de Saúde. 2007;16(4):233–244.

25 de Freitas Saldanha R, Bastos RR, Barcellos C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). Cadernos de Saúde Pública. 2019;35(9):1–9.

26 Honaker J, King G, Blackwell M. Amelia II: A Program for Missing Data. Journal of Statistical Software. 2011;45(7):1–47.

27 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2013.

28 Kleinbaum DG, Klein M. Logistic Regression. 1st ed. New York: Springer; 2010.

29 Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B. 1996;58(1):267–288.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.12.24301163; this version posted January 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

30 Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005;67:301–320.

31 Breiman L. Random Forests. Machine Learning. 2001;45:5–32.

32 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: ACM; 2016. p. 785–794. Available from: `http://doi.acm.org/10.1145/2939672.2939785`.

33 Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks; 1984.

34 Aggarwal CC. Neural Networks and Deep Learning: A Textbook. 1st ed. Cham: Springer; 2018.

35 Miao C, Bao M, Xing A, Chen S, Wu Y, Cai J, et al. Cardiovascular Health Score and the Risk of Cardiovascular Diseases. PLOS ONE. 2015;10(7):e0131537.

36 Li R, Chen Y, Ritchie MD, Moore JH. Electronic health records and polygenic risk scores for predicting disease risk. Nature Reviews Genetics. 2020;21:493–502.

37 Szwarcwald CL, de Andrade CLT, Bastos FI. Income inequality, residential poverty clustering and infant mortality: a study in Rio de Janeiro, Brazil. Social Science & Medicine. 2002;55(12):2083–2092.