Overachieving Municipalities in Public Health: A Machine-learning Approach

Alexandre Dias Porto Chiavegatto Filho,[a,b] *Hellen Geremias dos Santos,*[a] *Carla Ferreira do Nascimento,*[a] *Kaio Massa,*[a] *and Ichiro Kawachi*[b]

Background: Identifying successful public health ideas and practices is a difficult challenge towing to the presence of complex baseline characteristics that can affect health outcomes. We propose the use of machine learning algorithms to predict life expectancy at birth, and then compare health-related characteristics of the under- and overachievers (i.e., municipalities that have a worse and better outcome than predicted, respectively).

Methods: Our outcome was life expectancy at birth for Brazilian municipalities, and we used as predictors 60 local characteristics that are not directly controlled by public health officials (e.g., socioeconomic factors).

Results: The highest predictive performance was achieved by an ensemble of machine learning algorithms (cross-validated mean squared error of 0.168), including a 35% gain in comparison with standard decision trees. Overachievers presented better results regarding primary health care, such as higher coverage of the massive multidisciplinary program Family Health Strategy. On the other hand, underachievers performed more cesarean deliveries and mammographies and had more life-support health equipment.

Conclusions: The findings suggest that analyzing the predicted value of a health outcome may bring insights about good public health practices.

Keywords: Brazil, Life expectancy, Machine learning, Prediction

(*Epidemiology* 2018;29: 836–840)

Machine learning algorithms have been successfully employed for complex health-related problems such as disease diagnosis,[1] mortality risk prediction,[2] and evaluating adverse birth risks,[3] but its potential in public health remains underexplored.

One of the most difficult challenges in public health is to be able to identify successful ideas and practices. Favorable health situations at the local level are not necessarily a consequence of good public health management, as there are a large number of characteristics with potentially complex interactions that can affect local health outcomes (e.g., municipalities with very high per capita income usually have high life expectancy, independently of the quality of their public health management).[4] We propose a machine learning framework to identify successful public health practices at the local level. We start by training algorithms to predict the value of a municipality's life expectancy at birth using as predictors only local characteristics that are not directly controlled by local public health managers (e.g., socioeconomic factors). Given a high predictive performance of the algorithms, we will then identify the municipalities that have a better outcome than predicted ("overachievers") and compare their health-related characteristics with the ones that have a worse outcome than predicted ("underachievers").

METHODS

We analyzed official public data from municipalities of Brazil, a country with both a recent history of reliable health data,[5] and the presence of large socioeconomic disparities that can affect health outcomes.[6,7] For our analyses, we included only the municipalities with more than 10,000 residents to decrease the influence of random annual variations in health outcomes (n = 3,052 from a total of 5,565 municipalities). All the data collected are from 2010, the year of the last Census, and the unit of analysis was each municipality. Our public health outcome of interest was life expectancy at birth for each municipality, calculated previously by the United Nations' Development Programme.[8] We used as predictors 60 local characteristics that are not directly controlled by public health officials and are publicly available at the municipal level, such as per capita income, proportion of illiterate residents, proportion of households that have computers, proportion of women, unemployment rate, proportion of married residents, proportion of white residents, and the state where the municipality is situated, among others (see eAppendix 1; http://links.lww.com/EDE/B406).

We tested the performance of a total of 16 widely available machine learning algorithms and then added an ensemble algorithm that creates an optimal weighted average of the

Submitted February 19, 2018; accepted September 6, 2018.

From the [a]Department of Epidemiology, School of Public Health of the University of Sao Paulo, Sao Paulo, SP, Brazil; and [b]Department of Social and Behavioral Sciences, Harvard T. H. Chan School of Public Health, Boston, MA.

Data and code are available in eAppendix9.

Supported by a grant from the Lemann Foundation (Harvard Brazil Research Fund) and FAPESP (grant number: 17/09369-8).

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Alexandre D. P. Chiavegatto Filho, Department of Epidemiology, School of Public Health, University of Sao Paulo, 715 Av Dr Arnaldo, Sao Paulo, SP, Brazil 01246-904. E-mail: alexdiasporto@usp.br.

Ma and Lin have contributed equally to this article.

ISSN: 1044-3983/18/301-0836
DOI: 10.1097/EDE.0000000000000919

individual models (SuperLearner), to predict life expectancy at birth. The algorithms included artificial neural networks, random forests, gradient boosted trees, least squares, and penalized linear regression (ridge and lasso), support vector machines, among others (see eAppendix 2; http://links.lww.com/EDE/B406).[9] Continuous variables were normalized to avoid oversized effects owing to difference in scales, and the categorical variable (the state where the municipality is situated) was transformed into indicator variables.

We then applied nested 10-fold cross-validation, composed by an outer and an inner validation process, to select the hyperparameters and weights of the different machine learning algorithm that composed the SuperLearner, and to test its performance.[10,11] The use of nested cross-validation performance allows the SuperLearner to be a good estimate of the expected error when tested on new (unseen) data.[11] The predictive performance of the algorithms was measured by the mean squared error. All the analyzes were performed with R software and model training with *caret* and *SuperLearner* package.[9] The study only used publicly available aggregated data and did not require ethical review.

Health Characteristics

Municipalities were separated into tertiles according to life expectancy at birth to compare municipalities with similar overall socioeconomic characteristics, where the first tertile is the one with lower life expectancy. Within each group, we ranked the under and overachievers based on their performance in life expectancy at birth (i.e., lower or higher than their predicted values). For each tertile, we selected the top 100 overachievers and top 100 underachievers and compared the health characteristics of the two groups using Mann-Whitney U tests, owing to non-normality of the distributions. Health characteristics analyzed included vaccination coverage, hospital beds per 10,000 residents, coverage of Estratégia Saúde da Família (Brazil's main primary health program),[12] ultrasound machines per 10,000 live births, among others (see eAppendix 3; http://links.lww.com/EDE/B406).

RESULTS

Life expectancy at birth for the 3,052 municipalities ranged from 65.5 years in Joaquim Nabuco, Pernambuco, to 78.6 years in Brusque, Santa Catarina (total range of 13.1 years). The best performing single algorithm, random forests, presented a cross-validated mean squared error of 0.175 (Figure 1). For random forests, residing in Minas Gerais State, the illiteracy rate and the proportion of households with an automobile were the most important variables for improving overall predictive performance (see eAppendix 4; http://links.lww.com/EDE/B406).

We then combined the original algorithms, with weights defined by 10-fold cross validation, to create the

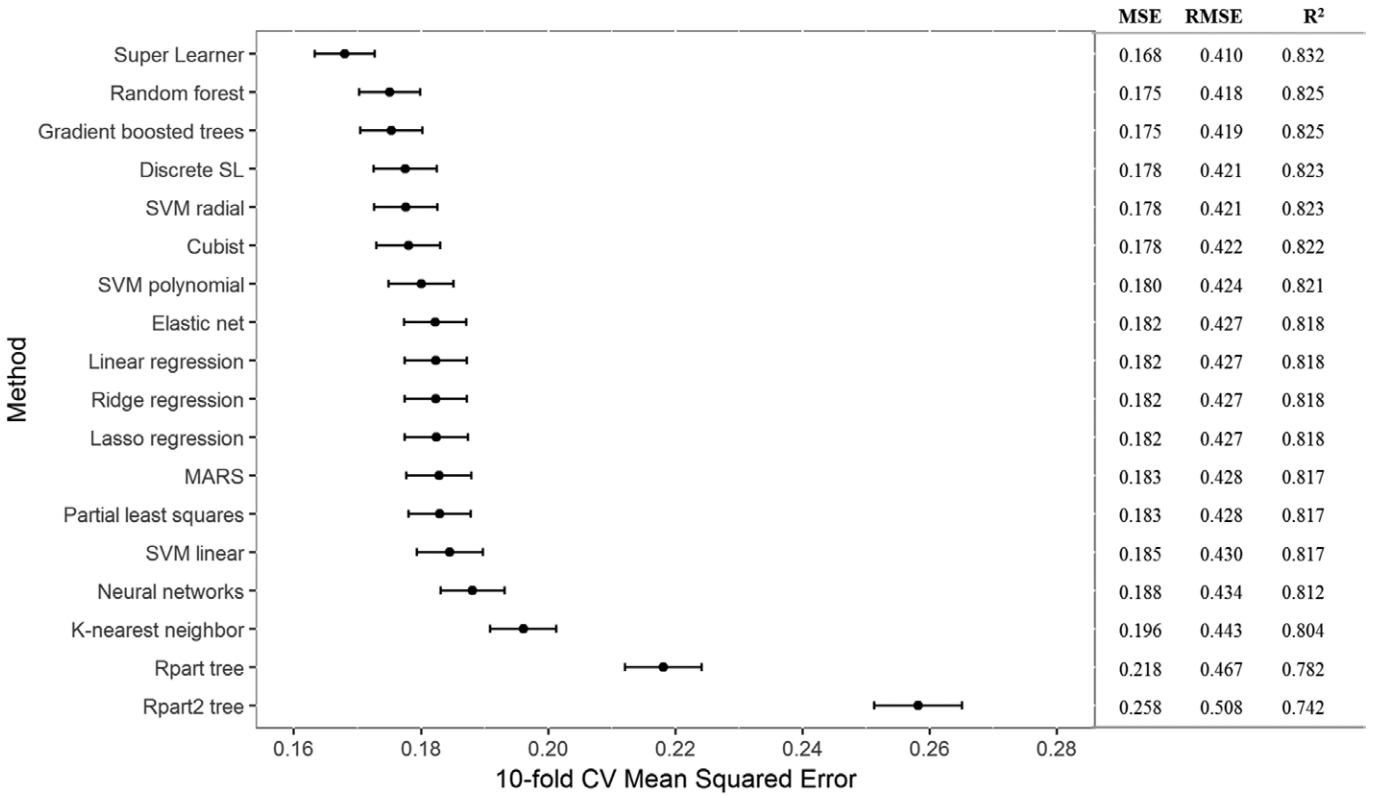

FIGURE 1. Cross-validated performance of the machine learning algorithms according to mean squared error (MSE), root mean squared error (RMSE) and R^2.

SuperLearner, which presented the best predictive performance (mean squared error: 0.168). The ensemble was used to make the final predictions on life expectancy at birth for each of the 3,052 municipalities and to identify the under and overachievers (eAppendix 5; http://links.lww.com/EDE/B406). The cross-validated values of predicted and actual life expectancies for each of the municipalities is presented in Figure 2. Overall, the algorithm performed well on the entire distribution, that is, when it predicted a high life expectancy the municipality presented a high life expectancy, and vice-versa. The map of the spatial distribution of the under and overachievers according to tertiles of life expectancy is presented in eAppendix 6; http://links.lww.com/EDE/B406. It is possible to observe an interestingly broad distribution of under and overachievers throughout Brazil without clear indications of clustering.

Table 1 presents the results for the differences in health-related characteristics between the under and overachievers. There were considerable differences between the two groups mostly for the middle and high tertiles of life expectancy. Overall, overachievers presented better results regarding investments in primary health care, that is, care usually provided by generalist physicians, dentists, and nurses, such as higher coverage of the massive multidisciplinary program Estratégia Saúde da Família (Family Health Strategy) and basic oral health coverage, and vaccination coverage. On the other hand, underachievers performed more cesarean deliveries and mammographies and had more life-support health equipment, which are usually provided by medical specialists (secondary health care).

DISCUSSION

Our results show that it is possible to predict with good accuracy the life expectancy of Brazilian municipalities using solely socioeconomic and demographic characteristics. We then analyzed health characteristics of the underachievers and overachievers, that is, municipalities that performed worse and better than predicted, given their socioeconomic characteristics, respectively, and found that overachievers performed better on primary health indicators, while underachievers fared better on secondary care indicators.

The results confirm the importance of primary healthcare investments identified in previous analyses that applied traditional epidemiologic methods,[13,14] highlighting the necessity of allocating resources to areas with higher marginal benefits, especially on a developing country with scarce resources like Brazil. Contrary to what could be expected, underachievers did not perform worse in all healthcare indicators, but in fact showed better results regarding medical technologies as evidenced by mammographies and cesarean deliveries performed, and the availability of life-support equipment.

The results also bring new knowledge to the healthcare literature regarding the growing area of machine learning. First, tree-based ensemble algorithms (random forests and gradient boosted trees) achieved the highest performance of the individual algorithms. Many recent articles resort to only testing one of the two for prediction problems, instead of comparing the performance of a broader group of algorithms, which our results show is an acceptable strategy. Second, an ensemble of machine learning algorithms (SuperLearner)

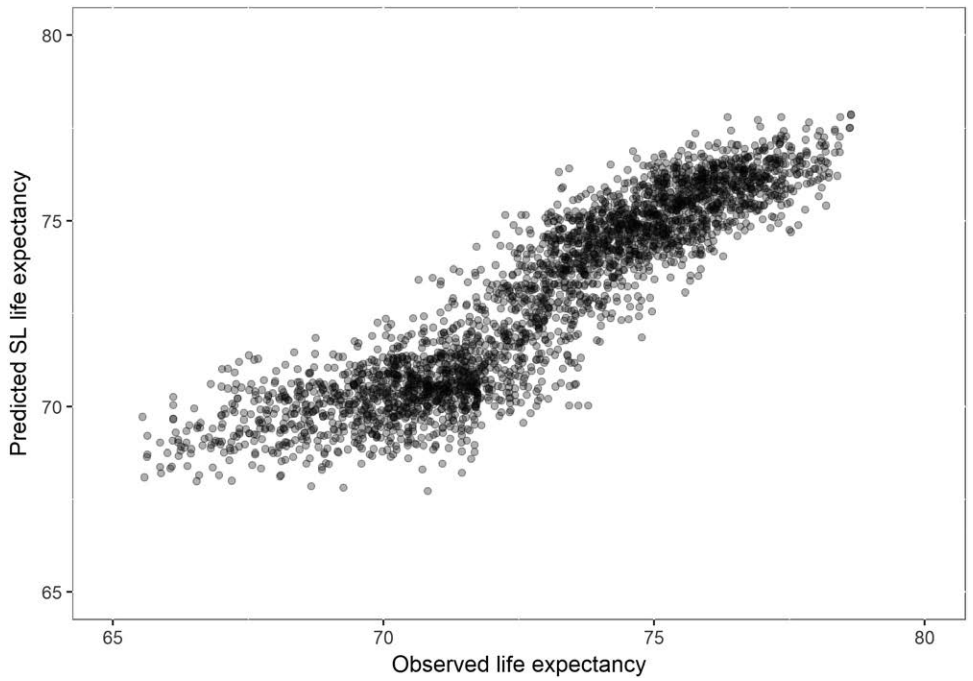

FIGURE 2. Cross-validated results for predicted and actual life expectancy at birth for the 3,052 municipalities.

TABLE 1. Differences in Health-related Characteristics Between Under and Overachievers, According to Tertiles of Life Expectancy at Birth

	Lower Tertile				Middle Tertile				Higher Tertile			
Healthcare Variables	**Over**	**Under**	**Difference[a]**	**CI[b] (95%)**	**Over**	**Under**	**Difference[a]**	**CI[b] (95%)**	**Over**	**Under**	**Difference[a]**	**CI[b] (95%)**
Cesarean deliveries (%)	30.27	33.56	-3.59	-7.17;-0.04	37.08	59.68	-21.34	-25.23; -17.25	57.03	62.90	-6.41	-10.60; -2.35
Family Health Strategy teams per 10,000 residents	3.40	3.2	0.10	-0.20; 0.40	3.00	2.20	1.10	0.70; 1.40	1.95	1.40	0.40	0.10; 0.80
Hospital beds per 10,000 residents	12.90	14.75	-1.60	-4.60; 0.60	15.00	16.40	-2.02	-5.50; 0.70	16.00	15.45	0.80	-1.90; 3.70
Life support equipment per 10,000 residents	0.80	1.3	-0.50	-0.90; -0.10	1.45	3.65	-2.10	-2.60; -1.60	3.00	3.25	-0; 00	-0.60; 0.50
Low birth weight (%)	7.11	6.64	0.51	-0.03; 1.06	6.96	8.20	-0.97	-1.68; -0.36	8.40	8.70	-0.49	-1.08; 0.11
Mammographies per 100 women	2.00	3.00	-0.00	-1.00; 0.00	3.00	13.00	-9.00	-11.00; -8.00	11.50	14.00	-4.00	-6.00; -2.00
Oral health coverage	69.74	84.72	-0.00	-7.34; 0.00	81.23	36.78	28.86	19.64; 39.86	38.88	21.87	11.00	0.00; 21.30
Primary health coverage for poor residents (last year)	100.00	94.81	0.00	-0.00; 0.00	85.51	76.46	14.18	5.31; 19.70	63.43	54.77	8.90	0.00; 18.42
Ultrasound machines per 10,000 live births	22.10	33.95	-5.60	-14.40; 0.00	27.60	43.65	-14.50	-25.30; -0.00	39.15	38.30	0.50	-8.00; 11.60
Vaccination coverage (%)	81.23	78.6	2.08	-0.99; 5.09	77.48	74.97	2.82	0.53; 5.26	75.88	73.52	2.16	0.31; 3.95
X-ray machines per 10,000 residents	0.40	0.5	-0.00	-0.20; -0.00	0.65	1.00	-0.50	-0.70; -0.30	1.00	1.10	-0.10	-0.30; 0.10

[a]For Mann-Whitney U tests, the difference in this case is the median of the difference between a sample from the underachievers and a sample from the overachievers, not the difference of the median between the two groups.
[b]Confidence intervals.

presented the highest predictive performance, including a 35% gain in performance in comparison with standard decision trees.

Our study introduced a new strategy for identifying successful healthcare initiatives, which brings a new approach to the outcome score literature.[15] The same strategy can be applied for other healthcare problems, such as identifying hospitals with a lower than expected mortality rate (after predicting the rate using the characteristics of the patients), identifying companies with more occupational accidents than expected (after predicting this number using the characteristics of industry and workforce), among many other applications.

The study has a few limitations. First, the selection of different algorithms could lead to different results. However, we performed the same analyses with the highest performing individual algorithm (random forests) and found similar results (see eAppendix 7; http://links.lww.com/EDE/B406). Second, there could be other relevant predictive variables not accounted for in our analyses. We aimed at including every socioeconomic characteristic that could plausibly affect life expectancy and was available in Brazil, but other variables unaccounted for could have an effect on prediction. However, the effect of each individual variable on improving predictive performance was limited. We tested the results by excluding the first, second, and third single variable that most improved predictive performance and the difference found was small (eAppendix 8; http://links.lww.com/EDE/B406).

Recent advances in machine learning have the potential to improve how healthcare is provided and evaluated.[16,17] As healthcare costs continue to increase, new pressures arise to cut access or availability of care, which could be more properly resolved by improving the efficiency of healthcare investments and management. Our study shows that assessing the predicted value of a health outcome may bring new insights about good public health practices.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
2. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38:500–507.
3. Pan I, Nolan LB, Brown RR, et al. Machine learning for social services: a study of prenatal case management in Illinois. *Am J Public Health*. 2017;107:938–944.
4. Berman LF, Kawachi I. *Social Epidemiology.* New York, NY: Oxford University Press; 2014.
5. Queiroz BL, Freire FHMA, Gonzaga MR, Lima EEC. Completeness of death-count coverage and adult mortality (45q15) for Brazilian states from 1980 to 2010. *Rev Bras Epidemiol*. 2017; Suppl 01:21–33.
6. Victora C. Socioeconomic inequalities in health: reflections on the academic production from Brazil. *Int J Equity Health*. 2016;15:164.
7. Landmann-Szwarcwald C, Macinko J. A panorama of health inequalities in Brazil. *Int J Equity Health*. 2016;15:174.

8. United Nations' Development Programme. *Human Development Atlas in Brazil*. 2013. Available at: http://atlasbrasil.org.br/2013/en. Accessed February 10, 2018.
9. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med*. 2015;3:42–52.
10. Kennedy C. Guide to SuperLearner. Available at: https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html. Accessed June 25, 2018.
11. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY: Springer; 2013.
12. Rasella D, Harhay MO, Pamponet ML, Aquino R, Barreto ML. Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide analysis of longitudinal data. *BMJ*. 2014;349:g4014.
13. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. *Milbank Q*. 2005;83:457–502.
14. Starfield B, Shi L, Grover A, Macinko J. The effects of specialist supply on populations' health: assessing the evidence. *Health Aff (Millwood)*. 2005;Suppl Web Exclusives:W5–97.
15. Wyss R, Glynn RJ, Gagne JJ. A review of disease risk scores and their application in pharmacoepidemiology. *Curr Epidemiol Rep*. 2016;3:277–284.
16. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216–1219.
17. Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind and the future of medicine. *N Engl J Med*. 2017;377:1209–1211.