https://doi.org/10.1590/1980-549720210047

ESPECIAL ARTICLE / *ARTIGO ESPECIAL*

Estimating underdiagnosis of COVID-19 with nowcasting and machine learning

Estimativa do subdiagnóstico de COVID-19 utilizando machine learning e nowcasting

Leandro Pereira Garcia[I], André Vinícius Gonçalves[II,III], Matheus Pacheco Andrade[I], Lucas Alexandre Pedebôs[I], Ana Cristina Vidor[I], Roberto Zaina[II], Ana Luiza Curi Hallal[IV], Graziela de Luca Canto[IV], Jefferson Traebert[V], Gustavo Medeiros de Araújo[II], Fernanda Vargas Amaral[VI]

ABSTRACT: ***Objective:*** To analyze the underdiagnosis of COVID-19 through nowcasting with machine learning in a Southern Brazilian capital city. ***Methods:*** Observational ecological design and data from 3916 notified cases of COVID-19 from April 14th to June 2nd, 2020 in Florianópolis, Brazil. A machine-learning algorithm was used to classify cases that had no diagnosis, producing the nowcast. To analyze the underdiagnosis, the difference between data without nowcasting and the median of the nowcasted projections for the entire period and for the six days from the date of onset of symptoms were compared. ***Results:*** The number of new cases throughout the entire period without nowcasting was 389. With nowcasting, it was 694 (95%CI 496–897). During the six-day period, the number without nowcasting was 19 and 104 (95%CI 60–142) with nowcasting. The underdiagnosis was 37.29% in the entire period and 81.73% in the six-day period. The underdiagnosis was more critical in the six days from the date of onset of symptoms to diagnosis before the data collection than in the entire period. ***Conclusion:*** The use of nowcasting with machine learning techniques can help to estimate the number of new disease cases.

Keywords: COVID-19. Underregistration. Nowcast. Machine learning.

[I]Prefeitura de Florianópolis – Florianópolis (SC), Brazil.
[II]Information Sciences Center, Universidade Federal de Santa Catarina – Florianópolis (SC), Brazil.
[III]Instituto Federal do Norte de Minas Gerais – Montes Claros (MG), Brazil.
[IV]Health Sciences Center, Universidade Federal de Santa Catarina – Florianópolis (SC), Brazil.
[V]Post-Graduation Program in Health Sciences, Universidade do Sul de Santa Catarina – Palhoça (SC), Brazil.
[VI]Universidad de Málaga – Málaga, Spain.
Corresponding author: Jefferson Traebert. Avenida Pedra Branca, 25, Cidade Universitária Pedra Branca, CEP: 88132-270, Palhoça (SC), Brazil. E-mail: jefferson.traebert@gmail.com

Conflict of interests: nothing to declare – Financial support: none.

RESUMO: ***Objetivo:*** Analisar o subdiagnóstico da COVID-19 por meio de *nowcasting* com *machine learning* em uma capital do sul do Brasil. ***Métodos:*** Estudo ecológico observacional utilizando dados de 3.916 casos notificados de COVID-19 de 14 de abril a 2 de junho de 2020 em Florianópolis, Brasil. O algoritmo de *machine learning* foi usado para classificar os casos que ainda não tinham diagnóstico, produzindo o *nowcasting*. Para analisar o subdiagnóstico, foi comparada a diferença entre os dados sem *nowcasting* e a mediana das projeções com *nowcasting* para todo o período e para os seis dias a partir da data de início dos sintomas. ***Resultados:*** O número de novos casos sem *nowcasting* durante todo o período foi de 389, com *nowcasting* foi de 694 (IC95% 496–897). No período de seis dias, o número sem *nowcasting* foi de 19 e 104 (IC95% 60–142) com *nowcasting*. O subdiagnóstico foi de 37,29% em todo o período e 81,73% no período de seis dias. O subdiagnóstico foi mais crítico em seis dias, desde a data do início dos sintomas até o diagnóstico antes da coleta de dados, do que em todo o período. ***Conclusão:*** O uso de *nowcasting* com técnicas de *machine learning* pode ajudar a estimar o número de novos casos da doença.

Palavras-Chave: COVID-19. Sub-Registro. Prognóstico imediato. Aprendizado de máquina.

INTRODUCTION

The World Health Organization has reported more than 10 million cases of SARS-CoV-2 infection and 500,000 deaths[1], a significant part of which have occurred in Brazil. According to the Brazilian Ministry of Health, the country has reached more than 1,3 million cases and 58,000 deaths[2], which is aligned with the Imperial College London prediction of a higher number of deaths caused by COVID-19[3]. Brazil has the highest number of deaths among the Latin American countries[4]. The Lancet has dedicated, recently, an editorial to the political-sanitary disaster that soars the country[5]. Despite the already alarming numbers, the editorial[5] and other studies[6-11] have drawn attention to the possibility of a large number of underdiagnosed cases. One of the causes of underdiagnosis is the low testing rate in individuals with suspected cases: 4.71 tests for a thousand habitants[12]. This rate is much lower than in countries like Iceland (184.11), United States (66.76), Chile (30.01), and South Africa (16.34)[12]. Addressing underdiagnosis is essential so that appropriate actions can be taken to stop the progression in the number of deaths in the country[13].

Many countries are using a combination of containment and mitigation activities to stem the progression of SARS-CoV-2 and thus, manage the demand for hospital beds[14]. Non-pharmacological measures have been shown to be effective in controlling the transmission of COVID-19[15-19]. They can reduce the impact on the healthcare system, giving managers time to properly organize the system. These measures also reduce the need for hospitalization by other conditions that could compete for beds with SARS-CoV-2 infections[20]. In addition, they increase the chance that a substantial number of people avoid being infected until a treatment option and a vaccine are developed.

In outbreak situations, in which rapid changes are common, the actual number of infected cases must be closely monitored. Incidence variations produced by a suboptimal testing capacity should be distinguished from the real cases variation during the monitoring process[13]. If the number of individuals notified as suspects is much higher than the testing capacity at the present time, this difference can cause an underdiagnosis of the current cases. Data about pathogens transmissibility and exposed population susceptibility, population density, and demographic characteristics of the affected population, besides the temporal and spatial distribution of cases and population mobility, can contribute to the correction of such artefacts[21].

The natural history of the disease, on the other hand, is an important factor in determining the optimal case count update in frequency monitoring. Rapidly progressing diseases like COVID-19 require daily updates, while monthly updates may be sufficient for other diseases with slower progression, such as HIV/AIDS. A frequent analysis may also be necessary in times when transmissibility is expected to be changing, for example, when control actions are initiated, enhanced, or stopped[21].

Nowcasting approaches try to estimate the number of a given event in the present[13,15,22]. This strategy has been used to improve surveillance of infectious diseases like AIDS[23,24], cholera[25], influenza infections[13,26], and recently, COVID-19[3,8,17,20,22,27]. Nowcasting techniques, in general, use time-series predictions[28-30]. Recent advances in machine learning techniques offer opportunities to fine-tune epidemic behavior nowcasting[21]. The main objective of machine learning techniques is to produce a model that can be used to classify, predict, or estimate a phenomenon. This approach is useful in several applications in biomedical research[31-37], including concerning COVID-19[38,39].

Monitoring the impact of non-pharmacological actions is essential to optimize the allocation of scarce resources in non-high-income-countries, like Brazil[21]. In these countries, the maintenance of long quarantine periods is even more challenging due to a deficient social protection system, the economic vulnerability of the population, and the large portion of people acting as informal workers. No single set of interventions is appropriate to all contexts owing to the combination of these factors with the climatic, demographic, and organizational issues of each country[15]. Thus, monitoring in near real-time should be a key part of the strategy to couple with SARS-CoV2 infections. Among the challenges for timely monitoring are delays in providing medical care after the onset of symptoms and delays in diagnosis[13]. It is plausible to assume that these challenges are even greater in non-high-income-countries that have healthcare systems that are not widely available to the population.

To help overcome this challenge, the present study aimed to analyze the underdiagnosis of COVID-19 cases through nowcasting with machine learning in a Southern Brazilian capital city.

METHODS

ETHICAL CONSIDERATIONS

This project was submitted to the Human Research Ethics Committee at the Federal University of Santa Catarina to guarantee compliance with Resolution N°. 466/2012 of the National Health Council of Brazil. The research project was approved. We used secondary and anonymized databases only.

STUDY DESIGN

The present study has an observational ecological design, using data from notified cases of COVID-19 from the Health Department of Florianópolis, capital of the State of Santa Catarina, in southern Brazil, from April 14th to June 2nd, 2020. Florianópolis has 500,973 inhabitants[40] and is administratively divided into 49 health regions[41]. The health regions correspond to the areas covered by each primary health care unit. Detailed demographic data (total population and population by sex, education, income, race/skin color, and age) for each of the 49 health regions of Florianópolis were made available by the Health Department of Florianópolis and can be accessed at: https://github.com/lpgarcia18/underdiagnosis_of_covid_19_cases_in_brazil/tree/master/dados/demografia. The median time from the onset of COVID-19 symptoms to notification is three days; as well as the time from notification to test result availability in the city (unpublished data, provided by the Public Health Department of Florianópolis).

We used the random forest[42] machine learning algorithm to classify the notified cases which still did not have a diagnosis, producing the nowcast. To analyze the underdiagnosis, we compared the difference between data without nowcasting and the median of the nowcasted projections for the entire period of analysis and for the period from May 28th to June 2nd, 2020. The latter corresponds to the six days from the onset of symptoms to diagnosis at the moment of data extraction.

DEFINITION OF SUSPECTED AND CONFIRMED CASES

Notification of suspected cases of COVID-19 within 24 hours is mandatory in Brazil[38]. As of April 14th, 2020, Florianópolis adopted the same COVID-19 notification criteria as the Brazilian Ministry of Health: fever accompanied by cough, dyspnea, runny nose, or sore throat[43]. The cases have been confirmed by real-time reverse-transcriptase-polymerase-chain-reaction (RT-PCR), serological tests, or clinical-epidemiological criteria.

DATA SOURCE AND VARIABLES

Three data sources for the nowcasting were used, all from the Public Health Department of Florianópolis:

1. anonymized database of suspected and confirmed cases of Florianópolis residents;
2. demographic data for the 49 health regions; and
3. traffic data, as a proxy for the mobility of people in the municipality.

The following variables were extracted from an anonymized database of suspected and confirmed cases:

I. diagnostic (confirmed, ruled out, or missing),
II. sex,

III. age (in years),
IV. age groups (under 10 years, from 10 to under 20, from 20 to under 40, from 40 to under 60, from 60 to under 80 and over),
V. race (white and not),
VI. date of birth, and
VII. onset of symptoms.

The number of infected people (with a positive diagnosis and less than 14 days of symptom onset) and the rate of infections per 100,000 inhabitants were calculated for the health regions where each notified person resides. In addition, the following demographic data from these regions were included in the analysis:

I. the total number of inhabitants and the number by sex,
II. the number of persons aged 1 year old, 2 years old, and so on up to 100 years old or more,
III. the number of people by race (white, black, yellow, brown, indigenous, and ignored),
IV. the number of people by years of schooling (from 1–17 years completed or more, in addition to literate, non-literate, literate through youth and adult literacy programs, and uninformed schooling),
V. total income per household, average income of households, total income of heads of households, average income of heads of households, total income per person, and average income per person. The proportion of male people, people aged 60 years old or over, people of non-white race, and people with 10 or less schooling time was calculated as possible indicators of vulnerability.

The average daily traffic in four important avenues in the city was used as a proxy for people's mobility in the city. We assumed there is a lag between increased mobility and perception of an increase in the number of cases, so we used the average traffic of the day and the average daily traffic lagged until the thirteenth day of the onset of symptoms for the notified cases. There was no imputation for missing data.

DESCRIPTIVE ANALYSIS

To compare the characteristics of people with a confirmed and a ruled out diagnosis of COVID-19, t-test was used for continuous variables and χ^2 for categorical variables, adopting the $p<0.05$ as a threshold of statistical significance.

COVID-19 INCIDENCE NOWCASTING WITH RANDOM FOREST

We used the random forest to carry out the nowcasting. The database was initially divided into the training-validation-test database, formed by cases whose diagnosis (confirmed or discarded) was known; and the prediction database, which had no diagnosis. The training-validation-test database was then divided into a training-validation database and a test database, using 70 and 30% of the data, respectively.

The training-validation database was undersampled to improve the sample's balance as the number of cases ruled out was much higher than the number of cases confirmed. The undersampled training-validation database was used to perform feature selection and hyperparameter tuning in nested cross-validation with five folds in the inner and outer loops. Feature selection and hyperparametrization were performed simultaneously in the inner loop. The following hyperparameters and ranges were established: number of trees (100–2000), mtry (1–50), minimum node size (1–10), sample fraction (0–1), and percentage of features selected (0.2–1). Random search to select the hyperparameters seeking to maximize accuracy was used.

We analyzed the training and validation results. The model with the best fit was used for classification in the test database. The test database was not submitted to undersampling reflecting the prediction database as close as possible. Finally, the cases were classified as confirmed or discarded, based on the predictions.

We repeated the resampling of the databases, the training, and the testing of the algorithms 1000 times to determine the 95% Uncertainty Intervals (UI), the median of accuracy, sensitivity, and specificity, in addition to the final classification of cases.

The underdiagnosis was analyzed by the difference between the median of the number of cases predicted by the model (incidence with nowcasting) and the number of cases diagnosed by the Public Health Department of Florianópolis (incidence without nowcasting). This analysis was carried comparing the entire period and the period from May 28th to June 2nd, 2020. The number of cases was also smoothed by a LOESS[44] regression and the cumulative number, without and with nowcasting, were presented graphically by day of symptom onset.

All analyzes were performed using the software R v.3.6.3. Scripts and databases are available at: https://github.com/lpgarcia18/underdiagnosis_of_covid_19_cases_in_brazil

RESULTS

During the analysis period, 3916 individuals residing in Florianópolis were reported as suspected cases of COVID-19. Among all notified individuals, 603 had a positive diagnosis, 2413 had a negative diagnosis, and 900 still did not have a diagnosis. The association of individual characteristics, health regions, and mobility of people with confirmed or discarded cases can be seen in Table 1.

There was an apparent positive correlation between the mtry and the percentage of selected features. The mtry defines the number of variables randomly sampled as candidates in each split in the random forest. The percentage of selected features keeps a certain percentage of the most important variables. Thus, it is likely that after the random choice of variables performed by the mtry, the percentage of selected resources acted to cause a higher prevalence of high-importance variables.

Table 1. Association between individual characteristics, health territory characteristics, and social distancing and positive and negative cases of COVID-19 in Florianópolis, Santa Catarina, Brazil.

Features	Positive (n=603)	%	Negative (n=2413)	%	Total (n=3016)	%	p-value
Individual Notification Characteristics							
Average date of symptom onset (SD of date in days)	2020-04-22 (24.5)		2020-04-29 (17.4)		2020-04-28 (19.2)		<0.001
Average date of suspected case notification (SD of date in days)	2020-05-10 (13.7)		2020-05-06 (13.5)		2020-05-07 (13.6)		<0.001
Gender							
Female	306	50.7	1,369	56.7	1,675	55.5	0.008
Male	297	49.3	1,044	43.3	1,341	44.5	
Race/skin color							
Black	35	5.8	127	5.3	162	5.4	0.025
Brown	36	6.0	77	3.2	113	3.7	
White	503	83.4	2,079	86.2	2,582	85.6	
Yellow	29	4.8	129	5.3	158	5.2	
Missing	–	–	1	0.0	1	0.0	
Age							
Mean (SD)	40.7 (17.2)		37.6 (18.7)		38.2 (18.5)		<0.001
Age group							
Less than 10	17	2.8	214	8.9	231	7.7	<0.001
10–20	30	5.0	116	4.8	146	4.8	
20–40	264	43.8	1,039	43.1	1,303	43.2	
40–60	206	34.2	765	31.7	971	32.2	
60–80	74	12.3	226	9.4	300	9.9	
More than 80	12	2.0	53	2.2	65	2.2	
Health territory characteristics							
Population							
Mean (SD)	17,357.4 (10,731.0)		15,338.4 (8,993.6)		15,742.1 (9,399.7)		<0.001

Continue...

Table 1. Continuation.

Features	Positive (n=603)	%	Negative (n=2413)	%	Total (n=3016)	%	p-value
Male proportion							
Mean (SD)	0.92 (0.06)		0.93 (0.05)		0.93 (0.06)		<0.001
Infected people							
Mean (SD)	19.0 (8.6)		16.6 (10.1)		17.06 (9.9)		<0.001
Rate of infected people							
Mean (SD)	154.5 (137.1)		141.9 (146.5)		144.4 (144.7)		0.054
Percentage of people over 60 years of age							
Mean (SD)	0.16 (0.06)		0.14 (0.05)		0.14 (0.06)		<0.001
Percentage of people with less than 10 years of schooling time							
Mean (SD)	0.6 (0.06)		0.6 (0.06)		0.6 (0.06)		<0.001
Non-white race percentage							
Mean (SD)	0.16 (0.10)		0.17 (0.10)		0.16 (0.10)		0.045
Mean per capita income							
Mean (SD)	R$ 3,685.69 (R$ 1,898.76)		R$ 3,040.96 (R$ 1,609.68)		R$ 3,169.86 (R$ 1,690.92)		<0.001
Social distancing							
Traffic average lag 13 days							
Mean (SD)	12,217.4 (6,658.9)		10,462.3 (5,032.1)		10,813.2 (5,440.9)		<0.001

SD: standard deviation.

The group of individuals with a positive result for SARS-COV-2 had an earlier symptom onset date and later notification dates than individuals with negative results. There was also a difference regarding the distribution according to sex and race between the two groups. The average age among confirmed cases was higher than among cases ruled out. There was a heterogeneous distribution of confirmed and ruled out cases among the 49 health regions in the municipality. The average number of confirmed cases was higher in regions with a higher age average, proportion of women, average level of education, income, and white people. Most positive cases were observed after seven days of a higher car traffic average.

The classification algorithm showed an accuracy of 0.91 (95%CI 0.83–0.97) in the training database, 0.66 (95%CI 0.62–0.69) in the validation, and 0.66 (95%CI 0.62–0.69) in the test database. The algorithm's sensitivity was: 0.91 (95%CI 0.83–0.98) in the training database, 0.65 (95%CI 0.61–0.69) in the validation, and 0.65 (95%CI 0.57–0.79) in the test database.

The algorithm's specificity was: 0.91 (95%CI 0.82–0.97) in the training database, 0.66 (95%CI 0.61–0.70) in the validation, and 0.66 (95%CI 0.60–0.71) in the test database.

The incidence without nowcasting throughout the entire period was 389 new cases. With the nowcasting, it was 694 (95%CI 496–897). From May 28th to June 02nd, 2020, the incidence without nowcasting was 19 new cases and 104 (95%CI 60–142) with nowcasting. Thus, the underdiagnosis was 37.29% in the entire period and 81.73% in six days from the date of onset of symptoms to diagnosis at the moment of data extraction. The difference in the progression of new cases with and without nowcasting can be seen in Figure 1.

DISCUSSION

COVID-19 data analysis represents a challenge for statisticians and epidemiologists in non-high-income-countries due to the magnitude of underreporting[45]. Even so, the number

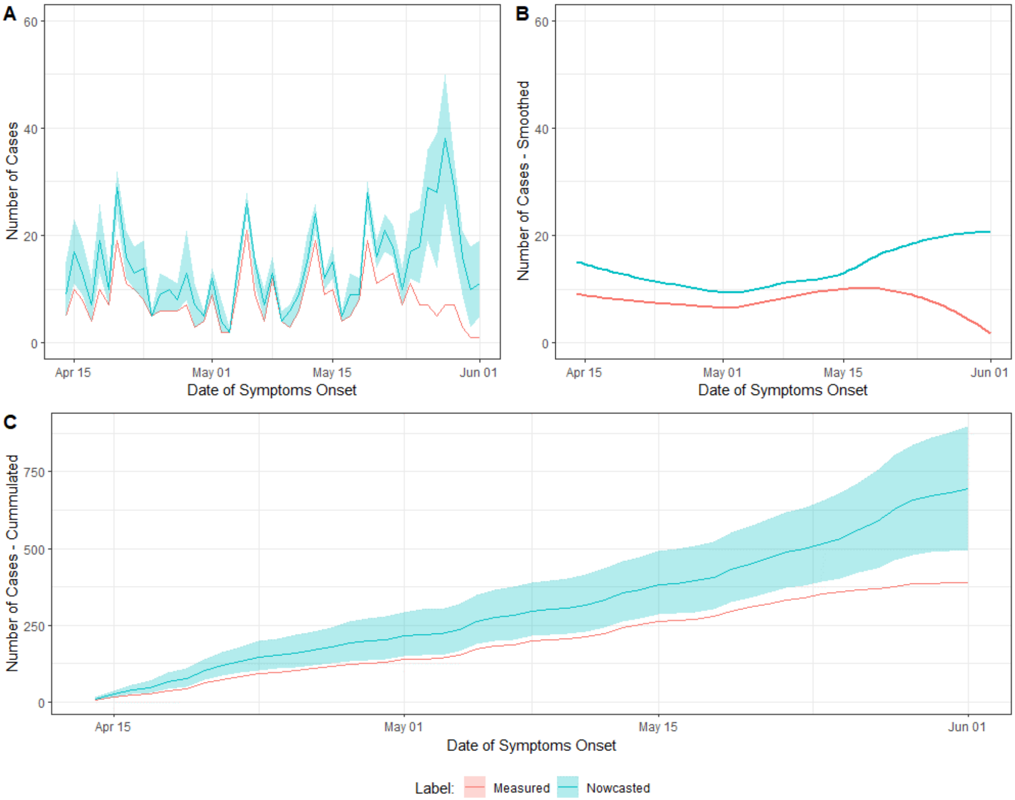

Figure 1. Evolution of the number of new cases of COVID-19 without nowcasting and with nowcasting. (A) Number of cases per day of symptom onset; (B) LOESS regression of the number of cases per day of symptom onset. (C) Cumulative number of cases per day of symptom onset. Key: the shaded area corresponds to the 95% uncertainty interval.

of COVID-19 cases has grown rapidly in Brazil, and today the country has the second largest number of cases in the world[12]. The city of Florianópolis had, so far, 1686 confirmed cases and 14 deaths caused by SARS-CoV-2[2]. At the time of this research, the epidemic in Florianópolis reflected the pattern of virus introduction in Latin America. Infections occurred first in people with higher income, who had traveled to countries where the virus was already present. Thus, a higher average of confirmed cases was observed in health regions with a greater number of white-skinned people, with higher income and schooling levels. Higher car traffic rates days before the onset of symptoms were also associated with confirmation of the reported cases. This corroborates evidence regarding the importance of social distancing to reduce the transmission of SARS-CoV-2. In a study carried out in Canada, for example, a social distancing strategy reduced the number of cases and the transmission of SARS-CoV-2, causing reduced intensive care units admissions and deaths[46].

Maintaining strict social distancing measures for long periods, however, may not be sustainable. These restrictions have already caused a slowdown in the world economy[47]. Research in non-high-income-countries shows an average 70% fall in income and a 30% decrease in consumption expenses[46]. Strategies that combine more restrictive periods with periods of relaxation have been identified as ideal for countries with few resources[48]. Alternating periods of greater social contact restriction of with periods of relaxation, but with an intensification of testing, isolation of people infected, contact tracing, and protection of vulnerable people, can allow people to have some social contact again economic production to be resumed[48].

Florianópolis has carried out more than 10,000 tests so far[50], which is more than 20 tests per a thousand inhabitants, more than four times the national average. Even with this greater number of tests, which should reduce the impact of underdiagnosis in the municipality, it is possible to observe a great disparity between the number of new cases confirmed by the municipality and the one predicted by the nowcasting. The underdiagnosis was more important in the proximal period of analysis. This shows how significant underdiagnosis is in the six days between the date of onset of symptoms and the date of diagnosis before data collection. Underdiagnosis, probably produced by a mismatch between the onset of symptoms and the time of testing, may interfere with the current estimate and future projections of disease incidence. In this context, the use of machine learning techniques can be helpful to enable adequate monitoring of the number of new cases and better decision making[50].

The algorithm performed better in detecting negative cases (specificity) than positive cases (sensitivity). In this sense, a greater number of false positives are expected compared to false negatives, and the interpretation of nowcasting should take this into account. A greater amount of individual data, such as data related to symptomatology, can improve model sensitivity. Besides, the association of SARS-CoV-2 infection rates with climate issues has been described[51,52]. The introduction of these data may also be useful and should be considered in future studies. The possible interaction between the mtry and the percentage of selected features needs to be better understood in future analysis. For future studies, other ensemble tree classifiers – such as XGBoost or Gradient Boosted Trees – could, also, be used as well as Bayesian models suitable for mixed data. Benchmarking could be used to evaluate the robustness of the chosen solution. Stationariness analyses of cases could bring pieces of evidence on the adequacy of the six-day window from the onset of symptoms and should be evaluated in other studies as well. In conclusion, the present study demonstrated an underdiagnosis of COVID-19 cases in Florianópolis. The underdiagnosis was more significant in the period of six days before data collection than in the entire period, corresponding to a monitoring artifact probably caused by a greater notification capacity compared to testing. An adequate and timely estimate of new cases is essential to monitor the number of reproductions and decision-making in the face of the epidemic. The use of nowcasting with machine learning techniques can help to estimate the number of new disease cases. A huge amount of new information about the COVID-19 pandemic is produced daily. Its use in improving health surveillance should be supported by the reformulation of prediction models as evidenced by this study.

REFERENCES

1. World Health Organization. Coronavirus disease (COVID-19) pandemic. Geneva: World Health Organization; 2020. [cited on Jun. 29, 2020]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=Cj0KCQjwoub3BRC6ARIsABGhnybzd7kDQxOQ-d5DH4OGL9618VaGon1x74u2OP0ujUw8vngt-huulrUaAsrqEALw_wcB
2. Brasil. Ministério da Saúde. Painel Coronavírus. Brasília: DATASUS; 2020. [cited on Jun. 29, 2020]. Available from: https://covid.saude.gov.br/
3. Bhatia S, Wardle J, Cori A, Parag KV, Mishra S, Cooper LV, et al. Short-term forecasts of COVID-19 deaths in multiple countries. London: Imperial College London; 2020. [cited on Jun. 29, 2020]. Available from: https://mrc-ide.github.io/covid19-short-term-forecasts/index.html
4. Simbana-Rivera K, Gomez-Barreno L, Guerrero J. Interim analysis of pandemic Coronavirus Disease 2019 (COVID-19) and the SARS-CoV-2 virus in Latin America and the Caribbean: morbidity, mortality and molecular testing trends in the region. medRxiv 2020: 20079863. https://doi.org/10.1101/2020.04.25.20079863
5. The Lancet. COVID-19 in Brazil: "So what?". Lancet 2020; 395 (10235): 1461. https://doi.org/10.1016/S0140-6736(20)31095-3
6. Reis RF, Quintela BM, Campos JO, Gomes JM, Rocha BM, Lobosco M, et al. Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. Chaos Solitons Fractals 2020; 136: 109888. https://doi.org/10.1016/j.chaos.2020.109888
7. Krantz SG, Rao ASRS. Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. Infect Control Hosp Epidemiol 2020; 41 (7): 857-9. https://doi.org/10.1017/ice.2020.116
8. Carvalho TA, Boschiero MN, Marson FAL. COVID-19 in Brazil: 150,000 deaths and the Brazilian underreporting. Diagn Microbiol Infect Dis 2021; 99 (3): 115258. https://doi.org/10.1016/j.diagmicrobio.2020.115258
9. Institute for Health Metrics and Evaluation. Estimation of total mortality due to COVID-19. Washington: IHME; 2021. [cited on Jun. 6, 2021]. Available from: http://www.healthdata.org/special-analysis/estimation-excess-mortality-due-covid-19-and-scalars-reported-covid-19-deaths
10. Quast T, Andel R. Excess mortality and potential undercounting of COVID-19 deaths by demographic group in Ohio. medRxiv 2020: 20141655. https://doi.org/10.1101/2020.06.28.20141655
11. Orellana JDY, Cunha GMD, Marrero L, Moreira RI, Leite IDC, Horta BL. Excess deaths during the COVID-19 pandemic: underreporting and regional inequalities in Brazil. Cad Saude Publica 2021; 37 (1): e00259120. https://doi.org/10.1590/0102-311X00259120
12. Worldometer. Covid-19 Coronavirus Pandemic 2020 [internet]. Worldometer; 2020. [cited on May 23, 2020]. Available from: https://www.worldometers.info/coronavirus/
13. McGough SF, Johansson MA, Lipsitch M, Menzies NA. Nowcasting by Bayesian Smoothing: a flexible,

generalizable model for real-time epidemic tracking. PLoS Comput Biol 2020; 16 (4): e1007735. https://doi.org/10.1371/journal.pcbi.1007735

14. Bedford J, Enria D, Giesecke J, Heymann DL, Ihekweazu C, Kobinger G, et al. COVID-19: towards controlling of a pandemic. Lancet 2020; 395 (10229): 1015-8.

15. Arslan S, Ozdemir MY, Ucar A. Nowcasting and forecasting the spread of COVID-19 and healthcare demand in Turkey: a modeling study. Front Public Health 2021; 8: 575145. https://doi.org/10.3389/fpubh.2020.575145

16. Ferguson N, Laydon D, Gilani GN, Imai N, Ainslie K, M Baguelin M, et al. Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. London: Imperial College London; 2020. [cited on May 30, 2020]. Available from: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/

17. Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. Lancet Public Health 2020; 5 (5): e279-88. https://doi.org/10.1016/S2468-2667(20)30090-6

18. Lai S, Ruktanonchai NW, Zhou L, Prosper O, Luo W, Floyd JR, et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. Nature 2020; 585 (7825): 410-3. https://doi.org/10.1038/s41586-020-2293-x

19. U.S. Department of Health and Human Services, Center for Disease Prevention and Control C. Implementation of mitigation strategies for communities with local COVID-19 transmission. 2020. [cited on Jul. 6, 2020]. Available from: https://stacks.cdc.gov/view/cdc/88478.

20. IHME COVID-19 health service utilization forecasting team, Murray JLC. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv 2020: 20043752. https://doi.org/10.1101/2020.03.27.20043752

21. Desai AN, Kraemer MUG, Bhatia S, Cori A, Nouvellet P, Herringer M, et al. Real-time epidemic forecasting: challenges and opportunities. Heal Secur 2019; 17 (4): 268-75. https://doi.org/10.1089/hs.2019.0022

22. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020; 395 (10225): 689-97. https://doi.org/10.1016/S0140-6736(20)30260-9

23. Cui J, Kaldor J. Changing pattern of delays in reporting AIDS diagnoses in Australia. Aust N Z J Public Health 1998; 22 (4): 432-5. https://doi.org/10.1111/j.1467-842x.1998.tb01409.x

24. Pagano M, Tu XM, Gruttola VD, MaWhinney S. Regression analysis of censored and truncated data: Estimating reporting-delay distributions and AIDS incidence from surveillance data. Biometrics 1994; 50 (4): 1203. PMID: 7787003

25. Pasetto D, Finger F, Camacho A, Grandesso F, Cohuet S, Lemaitre JC, et al. Near real-time forecasting for cholera decision making in Haiti after hurricane Matthew. PLoS Comput Biol 2018; 14 (5): e1006127. https://doi.org/10.1371/journal.pcbi.1006127

26. Spreco A, Eriksson O, Dahlström Ö, Cowling BJ, Timpka T. Evaluation of nowcasting for detecting and predicting local influenza epidemics, Sweden, 2009-2014. Emerg Infect Dis 2018; 24 (10): 1868-73. https://doi.org/10.3201/eid2410.171940

27. Jung S, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. Real-time estimation of the risk of death from novel Coronavirus (COVID-19) infection: Inference using exported cases. J Clin Med 2020; 9 (2): 523. https://doi.org/10.3390/jcm9020523

28. Bausch DG, Edmunds J. Real-time modeling should be routinely integrated into outbreak response. Am J Trop Med Hyg 2018; 98 (5): 1214-5. https://doi.org/10.4269/ajtmh.18-0150

29. Kassteele J, Eilers P, Wallinga J. Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing. Epidemiology 2019; 30 (5): 737-45. https://doi.org/10.1097/EDE.0000000000001050

30. Wu JT, Leung K, Leung GM. Review of "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020; 395 (10225): 689-97. https://doi.org/10.1016/S0140-6736(20)30260-9

31. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. The whole is greater than the sum of its parts: combining classical statistical and machine intelligence methods in medicine. Heart 2018; 104 (14): 1228. https://doi.org/10.1136/heartjnl-2018-313377

32. Fan H, Li L, Gilbert R, O'Callaghan F. A machine learning approach to identify cases of cerebral palsy using the UK primary care database. Lancet 2018; 392: S33. https://doi.org/10.1016/S0140-6736(18)32077-4

33. Wong D, Yip S. Machine learning classifies cancer. Nature 2018; 555 (7697): 446-7. https://doi.org/10.1038/d41586-018-02881-7

34. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature 2015; 521 (7553): 452-9. https://doi.org/10.1038/nature14541

35. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015; 349 (6245): 255-60. https://doi.org/10.1126/science.aaa8415
36. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018; 319 (13): 1317-8. https://doi.org/10.1001/jama.2017.18391
37. Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. JAMA Netw Open 2018; 1 (3): e180926. https://doi.org/10.1001/jamanetworkopen.2018.0926
38. Ribeiro MHDM, Silva RG, Mariani VC, Coelho LS. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Fractals 2020; 135: 109853. https://doi.org/10.1016/j.chaos.2020.109853
39. Chimmula VKR, Znang L. Time series forecasting of COVID-19 transmission in Canada using LSTM Networks. Chaos Solitons Fractals 2020; 135: 109864. https://doi.org/10.1016/j.chaos.2020.109864
40. Brasil. Instituto Brasileiro de Geografia e Estatística. Cidades. Rio de Janeiro: IBGE; 2020. [cited on Jun. 6, 2020]. Available from: https://cidades.ibge.gov.br
41. Brasil. Prefeitura de Florianópolis. Covidômetro. 2020. [cited on Jun. 29, 2020]. Available from: https://covidometrofloripa.com.br/
42. Breiman L. Random forests. Machine Learning 2001; 45: 5-32. https://doi.org/10.1023/A:1010933404324
43. Brasil. Ministério da Saúde. Emergência de Saúde Pública de Importância Nacional pela Doença pelo Coronavírus 2019. Vigilância integrada de síndromes respiratórias agudas doença pelo Coronavírus 2019, influenza e outros vírus respiratórios. 2020. [cited on May 29, 2020]. Available from: https://portaldeboaspraticas.iff.fiocruz.br/wp-content/uploads/2020/04/GuiaDeVigiEp-final.pdf
44. Jacoby WG. Loess: a nonparametric, graphical tool for depicting relationships between variables. Elect Stud 2000; 19 (4): 577-613. https://doi.org/10.1016/S0261-3794(99)00028-1
45. Hasell J, Ortiz-Ospina E, Mathieu E. Our world in data COVID-19 testing dataset. 2020. [cited on Jun. 30, 2020]. Available from: https://ourworldindata.org/coronavirus-data. Accessed June 30, 2020.
46. Tuite AR, Fisman DN, Greer AL. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. CMAJ 2020; 192 (19): E497-505. https://doi.org/10.1503/cmaj.200476
47. Baldé MAMT. Fitting SIR model to COVID-19 pandemic data and comparative forecasting with machine learning. medRxiv 2020: 20081042. https://doi.org/10.1101/2020.04.26.20081042
48. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Health 2020; 5 (5): e261-70. https://doi.org/10.1016/S2468-2667(20)30073-6
49. Florianópolis. Secretaria Municipal de Saúde. Coronavírus: Florianópolis testa 1,5 vezes mais que a Coréia do Sul. 2020. [cited on Jun. 30, 2020]. Available from: http://www.pmf.sc.gov.br/entidades/saude/?pagina=notpagina&menu=&noti=22422
50. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? Lancet 2020; 395 (10231): 1225-8. https://doi.org/10.1016/S0140-6736(20)30627-9
51. Gupta S, Raghuwanshi GS, Chanda A. Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. Sci Total Environ 2020; 728: 138860. https://doi.org/10.1016/j.scitotenv.2020.138860
52. Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A. Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. JAMA Netw Open 2020; 3 (6): e2011834. PMID: 32714105

Received on: 12/17/2020
Reviewed on: 07/12/2021
Accepted on: 08/02/2021
Preprint on: 07/02/2020
(https://doi.org/10.1101/2020.07.01.20144402)

Author's contributions: LPG: conceptualization, data curation, formal analysis, writing – original draft. AVG: conceptualization, data curation, formal analysis, writing – review & editing. MPA: conceptualization, data curation, formal analysis, writing – review & editing. LAP: conceptualization, data curation, formal analysis, writing – review & editing. ACV: conceptualization, data curation, formal analysis, writing – review & editing. RZ: conceptualization, data curation, formal analysis, writing – review & editing. ALCH: conceptualization, data curation, formal analysis, writing – review & editing. GLC: conceptualization, data curation, formal analysis, writing – review & editing. JT: conceptualization, data curation, formal analysis, writing – review & editing. GMA: conceptualization, data curation, formal analysis, writing – review & editing. FVA: conceptualization, data curation, formal analysis, writing – review & editing.