VOLUME 26 NO 9 PP 1019–1028 SEPTEMBER 2021

# How many more? Under-reporting of the COVID-19 deaths in Brazil in 2020

**Emil Kupek**

*Department of Public Health, Federal University of Santa Catarina, Florianopolis, Brazil*

**Abstract**

OBJECTIVE To evaluate the magnitude of under-reporting the number of deaths due to COVID-19 in Brazil in 2020, previously shown to occur due to low rate of laboratory testing for SARS-CoV-2, reporting delay, inadequate access to medical care, and its poor quality, leading to the low sensitivity of epidemiological surveillance and poor outcomes, often without laboratory confirmation of the cause of death.

METHODS Excess mortality due to COVID-19 was estimated directly based on various data sources, and indirectly, based on the difference between the observed and expected number of deaths from serious acute respiratory infection (SARI) and all-natural causes in 2020 had there been no COVID-19. The absence of laboratory testing for SARS-CoV-2 was adjusted based on the proportion of those who tested positive among the tested individuals whose death was attributed to COVID-19. Least absolute shrinkage and selection operator (lasso) were used to improve prediction of likely mortality without COVID-19 in 2020.

RESULTS Under-reporting of COVID-19 deaths was 22.62%, with a corresponding mortality rate per 100 000 inhabitants of 115 by the direct method, 71–76 by the indirect methods based on the excess SARI mortality and 95–104 by excess mortality due to natural causes. COVID-19 was the third cause of mortality that contributed directly with 18%, and indirectly with additional 10–11% to all deaths in Brazil in 2020.

CONCLUSIONS Underestimation of COVID-19 mortality between 1:5 and 1:4 is likely its lower bound. Timely and accurate surveillance of death causes is of the essence to evaluate the COVID-19 burden.

**keywords** Brazil, causes of death, COVID-19, mortality, SARI, underreporting

**Sustainable Development Goals:** Good Health and well-being

## Introduction

The first deaths from Coronavirus disease 2019 (COVID-19) in Brazil were reported in March 2020, followed by a steady rise and prolonged plateau, then by slow decline towards the end of the year, only to give way to the second wave in the last weeks of 2020 [1]. Despite considerable local variations in timing and intensity of the epidemic, as well as the type and duration of the mitigation efforts, it was clear by the end of 2020 that the country was approaching 200 000 deaths related to COVID-19, second only to the USA [2].

Monitoring of severe acute coronavirus 2 (SARS-CoV-2) infection and COVID-19 was included within an already existing epidemiological surveillance of respiratory viral agents [3], set up in 2009 by the Brazilian Ministry of Health because of the influenza A(H1N1) pandemic. In parallel, death certificate information on the natural causes of death has been assembled by a non-governmental organisation (NGO) [4], and federal states release the statistics on COVID-19 on their websites. All these sources have provided daily updates and summaries regarding important epidemiological characteristics. Brazilian Ministry of Health also publishes its special bulletin on COVID-19 and serious acute respiratory infections (SARI) on regular basis [5].

Media coverage and much of the scientific debate in Brazil have used reported numbers of COVID-19 deaths and cases, despite obvious limitations of such data for comparing states of different population sizes. Also, the impact of the sources of under-notification was rarely included in most of the estimates presented. This paper aims to compare the results of various statistical methods and data sources to evaluate the under-reporting of the

COVID-19 deaths in Brazil in 2020 and related mortality rates.

## Material and methods

Excess mortality due to COVID-19 was estimated directly based on various data sources on its mortality, and indirectly, based on the difference between the observed and expected number of deaths from SARI and all-natural causes in 2020. Three important methodological issues were addressed in this work: delay in and absence of laboratory testing for SARS-CoV-2 among those who died of SARI, estimation of the expected number of SARI deaths in the absence of COVID-19 epidemic in 2020, and accounting for confounding in estimating the time trend before COVID-19 epidemic in 2020.

### Direct method: adjustment for the lack of laboratory testing

Reverse transcription-polymerase chain reaction (RT-PCR) and serological testing are essential in confirming COVID-19 as a cause of death. However, in Brazil, many health facilities are ill-equipped to collect the specimen for such testing, and the delay in receiving the test results may pass the date of death. Nevertheless, the proportion of positive test results ($p_+$) among those tested for SARS-CoV-2 can be multiplied by the number of patients awaiting their test results ($N_a$) or without a chance of having this result ($N_w$), and summed up to those with positive test result ($N_p$) to estimate the total number of deaths from COVID-19 ($N_{tot}$):

$$N_{tot} = N_p + \left[p_+(N_a + N_w)\right] \qquad (1)$$

All individuals considered here are those who died, so they all had severe disease, and case severity was unlikely to bias the probability of SARS-CoV-2 testing.

### Indirect methods: estimation of the number of expected deaths

Another issue is the number of deaths that would have occurred had there been no COVID-19 epidemic – a hypothetical value known as a potential outcome or counterfactual [6]. Time series extrapolation and forecasting have been the most popular approaches but these do not account for (often unknown) confounding variables. To address this issue, the synthetic cohort (SC) method has been proposed and implemented in econometric end epidemiological studies [7–10]. In a regression trend analysis, substantive requirements of the SC method are no causal relationship between control cohorts and the outcome of interest, a stable temporal relationship between the control cohorts and the outcome of interest, and the predictive value of the cohorts to the outcome [10]. However, COVID-19 has profoundly affected the whole health system, such as hospital admissions from other causes, availability of health professionals and healthcare supplies. It is therefore difficult to imagine that any cohort characteristic remained stable regarding an outcome related to COVID-19, such as SARI, and the SC method assumptions are likely violated [11].

In the present study, inferential lasso (least absolute shrinkage and selection operator) [12] was used to adjust for confounding in estimating the time trend prior to the COVID-19 epidemic. Lasso selection of control cohorts is more robust against the violation of the SC method requirements because its optimisation method selects strong and stable trend predictors, especially with cross-validation [13,14]. This method was applied to the 2009-2019 annual data to estimate a likely number of SARI deaths in 2020, against which the observed SARI mortality in the same year should be compared.

Three estimation methods were used to predict the likely number of deaths from SARI and natural causes: double-exponential moving averages (DEMA), linear regression on the natural logarithm of the number of deaths, hereafter called log-normal model (LNM) and Poisson regression. DEMA was chosen to forecast one year because exponential distribution covers a wide range of non-linear models and gives more weight to more recent values [15]. Poisson and linear regression on the natural logarithm of the SARI deaths were applied to predict one year ahead with and without lasso adjustment.

Control variables used to adjust for confounding in inferential lasso were hospitalisation rates per 10 000 inhabitants over the 2009–2019 period for the following chapters of the tenth revision of the International Classification of Diseases (ICD-10): I-VI (Certain infectious and parasitic diseases; Neoplasms; Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism; Endocrine, nutritional and metabolic diseases; XI-XVII (Diseases of the digestive system; Diseases of the skin and subcutaneous tissue; Diseases of the musculoskeletal system and connective tissue; Diseases of the genitourinary system; Pregnancy, childbirth, and the puerperium; Certain conditions originating in the perinatal period; Congenital malformations, deformations and chromosomal abnormalities) [16]. The main criterion for the control variable selection was their statistical independence regarding the outcome. For example, the hospitalisation rates due to respiratory diseases and those of the circulatory system did not fulfil this criterion because moderate and severe COVID-19 manifestations are primarily associated with these symptoms.

To calculate the impact of COVID-19 on mortality on a yearly basis, it is necessary to adjust for the actual duration of the epidemic in 2020. In Brazil, this period was ten months, so the formula below puts it on the same scale as other annual statistics used and provides the risk of death attributable to COVID-19 in 2020:

$$\mathrm{AR} = \{[\mathrm{EMRR}(12/10) - 1]/\mathrm{EMRR}(12/10)\}100 \qquad (2)$$

where AR is the attributable risk and EMRR stands for an excess mortality rate ratio of observed versus expected deaths in 2020. This is just a variation of a well-known formula AR = (RR–1)/RR, only EMRR is a monthly average over ten epidemic months, multiplied by 12 to get per year basis.

The COVID-19 case definition followed the Brazilian Ministry of Health criteria: RT-PCR and serological testing, based on both clinical and epidemiological criteria, both clinical and medical imaging (X-ray, tomography), and only clinical criteria [5]. The case definition of SARI used in the present study was that of the Brazilian Ministry of Health: body temperature >37.8°C, and breathing difficulty or dyspnoea or $O_2$ saturation <95% in blood, and cough or sore throat, and the need for hospitalisation or death after having presented the aforementioned symptoms [3]. Natural causes of death include all ICD-10 chapters except chapter XX (external causes).

### Data sources

Only primary cause of death was available from the data sources. The number of deaths from COVID-19 and SARI, as well as the data necessary to correct the total number of deaths from COVID-19 ($N_{\mathrm{tot}}$), was extracted from tables 7 and 11 in a specialised bulletin [5]. The causes of death provided by the NGO [4] were grouped as COVID-19, SARI (included COVID-19), all respiratory, sepsis, all other and undetermined.

The federal states' data on COVID-19 and SARI are assembled on the OpenSUS website maintained by the federal government [17]. However, no data cleaning such as eliminating duplicates and inconsistent records is provided. The present study excluded duplicates and some records with inconsistencies in the order of dates (birth, first symptoms, hospitalisation and death).

Population data were taken from the Brazilian Institute of Geography and Statistics (acronym IBGE) [18].

All data were aggregated at the state level. Stata software [19] was used for all statistical analyses.

## Results

On average, 91.5% of the people whose death was attributed to COVID-19 were laboratory tested for SARS-CoV-2, and 92.93% of these were positive to RT-PCR or serological testing (Table 1). Multiplying these percentages indicates about 85% of COVID-19 deaths confirmed by laboratory testing. After correcting for the absence of testing data (see Formula 1 in the methods section), the Ministry of Health data produced the highest estimate of the number of COVID-19 deaths compared to the death certificates and the data reported by the federal states. The corrected data showed an average underestimation of 21.62%, with the range of 10.51–26.07% between the states (Table 1).

Among five estimation methods for the expected number of SARI deaths in 2020 had it been no COVID-19 epidemic, the estimates ranged between 83 873 with lasso LNM and 94 040 with Poisson regression (Table 2). The smallest root mean square error for these methods was 0.052 for the lasso LNM, so it was considered the best model and used in subsequent analysis.

In 2020, the average observed COVID-19 MR per 100 000 inhabitants was 115 (Table 3) and reached the highest values in the states of Rio de Janeiro (178), Amazonas (155), Federal District (153), Ceará (144), Pernambuco (137, São Paulo (137) and Roraima (134). COVID-19 was by far the largest cause of SARI deaths, as indicated by a considerable overlap between their mortality rates. Overall, about 3 of 4 (115/152 = 0.76) SARI deaths were due to COVID-19, with the range of 68-89% across the federal states. In terms of the relative risk of dying from SARI in 2020 compared to the expected (counterfactual) value in the absence of COVID-19, the former increased 2.25 (2.05, 2.46) times on average without and 2.08 (1.73, 2.51) times with lasso adjustment (Table 3).

By applying Formula 2, the excess SARI mortality attributable to COVID-19 was 63% (59%, 66%) for the unadjusted and 60% (52%, 67%) for the lasso-adjusted LNM estimates (bottom line in Table 3), with considerable variation between the states. These estimates translate into excess SARI MR of 96 and 91 per 100 000 (0.63 × 152 and 0.60 × 152) attributable to COVID-19 (Table 3).

LNM produced a similar excess of mortality from the natural causes in 2020 for both lasso-adjusted (18.4%) and unadjusted (16.1%) estimates (Table 4). This is equivalent to 28–29% of excess mortality due to COVID-19 regarding the natural causes (formula 2). Dividing the difference observed-expected by the

**Table 1** COVID-19 and SARI deaths in Brazil, 2020: Laboratory confirmation rate, unadjusted reports and adjustment for testing delay

| Sources | Brazilian Ministry of Health | | | | | Delay-adjusted | | Death certificates | | Federal states | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Federal states, district | LT (%) | LCR (%) | *N* deaths COVID-19 | *N* deaths investigated | *N* deaths no laboratory | *N* deaths COVID-19 | Bias (%) | *N* deaths COVID-19 | *N* deaths SARI | *N* deaths SARI | *N* deaths COVID-19 |
| Rondônia | 81 | 90 | 1723 | 298 | 9 | 1965 | −12.31 | 1675 | 39 | 1680 | 329 |
| Acre | 97 | 98 | 602 | 85 | 0 | 673 | −10.51 | 901 | 33 | 600 | 85 |
| Amazonas | 88 | 88 | 5415 | 1502 | 15 | 6560 | −17.46 | 3296 | 879 | 5409 | 1709 |
| Roraima | 72 | 73 | 634 | 122 | 2 | 733 | −13.51 | 768 | 18 | 614 | 120 |
| Pará | 91 | 93 | 7615 | 2690 | 56 | 9576 | −20.48 | 6093 | 977 | 7345 | 2783 |
| Amapá | 64 | 66 | 663 | 106 | 7 | 755 | −12.18 | 887 | 22 | 696 | 102 |
| Tocantins | 93 | 95 | 1195 | 264 | 6 | 1408 | −15.10 | 989 | 28 | 1215 | 419 |
| Maranhão | 87 | 90 | 3599 | 1289 | 27 | 4534 | −20.62 | 3102 | 971 | 3319 | 1266 |
| Piauí | 92 | 95 | 2390 | 595 | 68 | 2887 | −17.21 | 1858 | 92 | 2434 | 664 |
| Ceará | 95 | 98 | 10 538 | 3682 | 133 | 13 255 | −20.50 | 11 017 | 784 | 10 299 | 3858 |
| Rio Grande do Norte | 93 | 97 | 2373 | 834 | 111 | 3029 | −21.65 | 2561 | 274 | 2382 | 958 |
| Paraíba | 95 | 96 | 3754 | 1444 | 40 | 4786 | −21.57 | 3450 | 340 | 3677 | 1625 |
| Pernambuco | 99 | 100 | 10 008 | 4801 | 87 | 13 215 | −24.27 | 8602 | 5301 | 9750 | 4731 |
| Alagoas | 84 | 90 | 2612 | 966 | 31 | 3314 | −21.18 | 2657 | 355 | 2554 | 946 |
| Sergipe | 96 | 98 | 2566 | 366 | 2 | 2871 | −10.62 | 2369 | 93 | 2781 | 402 |
| Bahia | 92 | 96 | 8560 | 3577 | 68 | 11 042 | −22.47 | 9733 | 544 | 8483 | 3894 |
| Minas Gerais | 97 | 98 | 12 345 | 7440 | 283 | 16 978 | −27.29 | 16 044 | 1013 | 12 552 | 8408 |
| Espirito Santo | 97 | 98 | 3633 | 660 | 10 | 4172 | −12.92 | 6027 | 387 | 3603 | 676 |
| Rio de Janeiro | 72 | 73 | 25 851 | 4794 | 549 | 30 105 | −14.13 | 31 831 | 2232 | 26 946 | 5151 |
| São Paulo | 96 | 97 | 47 525 | 23 463 | 626 | 63 088 | −24.67 | 58 190 | 2708 | 48 363 | 28 317 |
| Paraná | 99 | 99 | 7747 | 4415 | 22 | 10 479 | −26.07 | 11 050 | 390 | 7982 | 5679 |
| Santa Catarina | 95 | 97 | 5227 | 1473 | 63 | 6378 | −18.05 | 6089 | 129 | 5166 | 1921 |
| Rio Grande do Sul | 97 | 97 | 9054 | 4087 | 48 | 11 819 | −23.40 | 10 527 | 421 | 9166 | 4569 |
| Mato Grosso do Sul | 97 | 97 | 2442 | 917 | 18 | 3094 | −21.08 | 2715 | 97 | 2379 | 1082 |
| Mato Grosso | 87 | 90 | 2085 | 346 | 46 | 2400 | −13.13 | 3521 | 64 | 2008 | 342 |
| Goiás | 90 | 93 | 7238 | 2252 | 206 | 9010 | −19.67 | 8086 | 238 | 6676 | 2257 |
| Distrito Federal | 95 | 96 | 4140 | 1016 | 28 | 4941 | −16.22 | 4457 | 75 | 4502 | 1322 |
| Total | 91 | 93 | 191 552 | 73 494 | 2561 | 244 396 | −21.62 | 218 493 | 18 502 | 192 581 | 83 615 |

LT, Laboratory tested by RT-PCR and/or serological tests; LCR, Laboratory confirmation rate; N, Number of events; SARI, Serious Acute Respiratory Infection (excluding COVID-19).

population results in MR per 100 000 of 105 and 95, respectively.

The MR for natural causes increased about 1% per year over the 2009–2019 period but the 2020 increase significantly exceeded the expected upper bound (Figure 1).

The lasso coefficients for the hospitalisation rates in adjusting for SARI mortality trend by lasso LNM were the following: congenital malformations/deformations and chromosomal abnormalities (−1.42), diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (0.61), diseases of the musculoskeletal system and connective tissue (−0.23), neoplasms (0.15), endocrine, nutritional/metabolic diseases (−0.11), certain infectious and parasitic diseases (0.04), and certain conditions originating in the perinatal

**Table 2** Expected number of deaths due to serious acute respiratory infection (SARI) in Brazil in 2020, based on different estimation methods applied to 2009–2019 annual data on SARI deaths

| Federal states, | DEMA | | | LNM | | | lasso LNM | | | Poisson regression | | | lasso Poisson | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| district | N | LB | UB | N | LB | UB | N | LB | UB | N | LB | UB | N | LB | UB |
| Rondônia | 433 | 392 | 474 | 549 | 508 | 594 | 578 | 531 | 625 | 412 | 386 | 440 | 813 | 797 | 829 |
| Acre | 281 | 248 | 314 | 225 | 206 | 246 | 250 | 219 | 281 | 263 | 242 | 285 | 293 | 284 | 303 |
| Amazonas | 869 | 811 | 927 | 1444 | 1344 | 1550 | 1363 | 1291 | 1436 | 951 | 903 | 1001 | 1419 | 1398 | 1440 |
| Roraima | 150 | 126 | 174 | 131 | 119 | 145 | 113 | 93 | 134 | 149 | 134 | 165 | 39 | 35 | 42 |
| Pará | 2913 | 2807 | 3019 | 3318 | 3083 | 3571 | 3045 | 2937 | 3153 | 2493 | 2340 | 2655 | 3954 | 3919 | 3990 |
| Amapá | 219 | 190 | 248 | 218 | 199 | 238 | 240 | 209 | 270 | 194 | 181 | 208 | 230 | 222 | 239 |
| Tocantins | 276 | 243 | 309 | 461 | 426 | 499 | 514 | 469 | 558 | 319 | 285 | 356 | 736 | 721 | 751 |
| Maranhão | 1906 | 1820 | 1992 | 2650 | 2465 | 2848 | 2514 | 2415 | 2612 | 1625 | 1469 | 1797 | 3083 | 3052 | 3114 |
| Piauí | 1714 | 1633 | 1795 | 1051 | 978 | 1130 | 1089 | 1024 | 1154 | 1296 | 1152 | 1458 | 1308 | 1288 | 1329 |
| Ceará | 5865 | 5715 | 6015 | 3567 | 3312 | 3841 | 3403 | 3289 | 3518 | 4480 | 4093 | 4904 | 4664 | 4625 | 4703 |
| Rio Grande do Norte | 1649 | 1569 | 1729 | 1191 | 1109 | 1280 | 1337 | 1265 | 1409 | 1476 | 1377 | 1583 | 1666 | 1643 | 1689 |
| Paraíba | 1998 | 1910 | 2086 | 1387 | 1292 | 1490 | 1436 | 1362 | 1510 | 1745 | 1558 | 1955 | 1946 | 1921 | 1971 |
| Pernambuco | 3551 | 3434 | 3668 | 3782 | 3511 | 4075 | 3740 | 3620 | 3860 | 3344 | 3138 | 3563 | 5027 | 4987 | 5067 |
| Alagoas | 1288 | 1218 | 1358 | 1122 | 1044 | 1206 | 1149 | 1083 | 1215 | 1146 | 1062 | 1237 | 1351 | 1330 | 1372 |
| Sergipe | 710 | 658 | 762 | 724 | 672 | 780 | 783 | 729 | 838 | 728 | 659 | 804 | 1118 | 1100 | 1137 |
| Bahia | 3696 | 3577 | 3815 | 6600 | 6090 | 7152 | 6200 | 6046 | 6354 | 3482 | 3321 | 3650 | 8203 | 8151 | 8254 |
| Minas Gerais | 10 044 | 9848 | 10 240 | 9639 | 8848 | 10 502 | 9167 | 8979 | 9354 | 10 073 | 9668 | 10 494 | 12 106 | 12 043 | 12 168 |
| Espirito Santo | 1439 | 1365 | 1513 | 1403 | 1307 | 1507 | 1609 | 1531 | 1688 | 1484 | 1381 | 1594 | 2395 | 2367 | 2422 |
| Rio de Janeiro | 10 764 | 10 561 | 10 967 | 7314 | 6740 | 7937 | 6965 | 6801 | 7128 | 11 919 | 11 290 | 12 583 | 10 794 | 10 735 | 10 853 |
| São Paulo | 23 247 | 22 948 | 23 546 | 23 592 | 21 317 | 26 110 | 20 547 | 20 266 | 20 828 | 28 485 | 26 597 | 30 506 | 27 171 | 27 077 | 27 264 |
| Paraná | 4770 | 4635 | 4905 | 4663 | 4320 | 5033 | 5141 | 5000 | 5282 | 4742 | 4497 | 5000 | 5303 | 5262 | 5344 |
| Santa Catarina | 2452 | 2355 | 2549 | 2713 | 2524 | 2916 | 2860 | 2755 | 2965 | 2382 | 2262 | 2507 | 3015 | 2983 | 3046 |
| Rio Grande do Sul | 4274 | 4146 | 4402 | 4605 | 4267 | 4970 | 5028 | 4889 | 5167 | 5407 | 4968 | 5885 | 5619 | 5576 | 5661 |
| Mato Grosso do Sul | 1215 | 1147 | 1283 | 888 | 825 | 956 | 960 | 899 | 1021 | 1251 | 1172 | 1337 | 1111 | 1092 | 1130 |
| Mato Grosso | 799 | 744 | 854 | 1136 | 1057 | 1221 | 1155 | 1089 | 1222 | 953 | 879 | 1033 | 1381 | 1360 | 1402 |
| Goiás | 2333 | 2238 | 2428 | 2605 | 2423 | 2799 | 2564 | 2465 | 2663 | 2606 | 2466 | 2753 | 4357 | 4319 | 4394 |
| Distrito Federal | 482 | 439 | 525 | 1047 | 974 | 1126 | 1061 | 997 | 1125 | 675 | 594 | 767 | 1576 | 1554 | 1599 |
| Total | 87 774 | 87 193 | 88 355 | 91 524 | 86 785 | 96 522 | 106 336 | 105 697 | 106 976 | 93 241 | 88 040 | 98 749 | 110 437 | 110 249 | 110 625 |

N, Expected number of deaths; DEMA, double-exponential moving averages; LNM, Log-normal model: linear regression with log-transformed number of deaths as the outcome; lasso, least absolute shrinkage and selection operator; LB, Lower bound of the 95% confidence interval; UB, Upper bound of the 95% confidence interval.

**Table 3** Estimated impact of COVID-19 on the mortality due to serious acute respiratory infection in Brazil, 2020

| | | | Expected by log-normal regression | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Observed MR | | SARI, no lasso | | | SARI with lasso | | |
| Federal states, district | COVID-19 | SARI | EMRR | LB | UB | EMRR | LB | UB |
| Rondônia | 106 | 126 | 4.77 | 4.56 | 4.98 | 3.34 | 3.19 | 3.48 |
| Acre | 78 | 88 | 2.56 | 2.37 | 2.75 | 2.66 | 2.46 | 2.86 |
| Amazonas | 155 | 191 | 6.90 | 6.73 | 7.06 | 4.80 | 4.68 | 4.91 |
| Roraima | 134 | 158 | 4.92 | 4.56 | 5.28 | 6.37 | 5.91 | 6.84 |
| Pará | 111 | 143 | 3.84 | 3.76 | 3.92 | 3.12 | 3.06 | 3.18 |
| Amapá | 90 | 103 | 3.89 | 3.61 | 4.17 | 3.12 | 2.90 | 3.34 |
| Tocantins | 88 | 105 | 4.41 | 4.18 | 4.64 | 2.71 | 2.57 | 2.85 |
| Maranhão | 64 | 82 | 2.79 | 2.71 | 2.87 | 1.83 | 1.78 | 1.89 |
| Piauí | 89 | 111 | 2.23 | 2.15 | 2.31 | 2.65 | 2.55 | 2.75 |
| Ceará | 144 | 187 | 2.96 | 2.91 | 3.01 | 3.91 | 3.85 | 3.98 |
| Rio Grande do Norte | 84 | 111 | 2.05 | 1.98 | 2.13 | 2.30 | 2.21 | 2.38 |
| Paraíba | 117 | 154 | 2.74 | 2.66 | 2.82 | 3.30 | 3.21 | 3.40 |
| Pernambuco | 137 | 188 | 3.95 | 3.88 | 4.02 | 3.56 | 3.50 | 3.62 |
| Alagoas | 97 | 126 | 2.89 | 2.79 | 2.99 | 2.85 | 2.75 | 2.94 |
| Sergipe | 122 | 138 | 3.94 | 3.80 | 4.09 | 3.61 | 3.48 | 3.74 |
| Bahia | 71 | 95 | 3.17 | 3.11 | 3.23 | 1.81 | 1.77 | 1.84 |
| Minas Gerais | 79 | 116 | 1.69 | 1.66 | 1.71 | 1.86 | 1.83 | 1.89 |
| Espirito Santo | 101 | 118 | 2.81 | 2.73 | 2.90 | 2.63 | 2.55 | 2.71 |
| Rio de Janeiro | 178 | 210 | 2.53 | 2.50 | 2.55 | 4.33 | 4.28 | 4.38 |
| São Paulo | 137 | 190 | 2.21 | 2.20 | 2.23 | 3.10 | 3.07 | 3.12 |
| Paraná | 91 | 130 | 2.21 | 2.17 | 2.25 | 2.07 | 2.03 | 2.11 |
| Santa Catarina | 88 | 109 | 2.68 | 2.61 | 2.74 | 2.25 | 2.19 | 2.30 |
| Rio Grande do Sul | 104 | 140 | 2.19 | 2.15 | 2.23 | 2.40 | 2.36 | 2.45 |
| Mato Grosso do Sul | 110 | 145 | 2.47 | 2.39 | 2.56 | 3.24 | 3.13 | 3.35 |
| Mato Grosso | 69 | 81 | 2.52 | 2.42 | 2.62 | 2.07 | 1.99 | 2.16 |
| Goiás | 128 | 164 | 3.46 | 3.39 | 3.53 | 3.37 | 3.30 | 3.44 |
| Distrito Federal | 153 | 187 | 7.32 | 7.12 | 7.52 | 4.36 | 4.24 | 4.48 |
| Total | 115 | 152 | 2.79 | 2.73 | 2.85 | 2.90 | 2.84 | 2.95 |

MR, Mortality rate per 100 000 inhabitants; SARI, Serious Acute Respiratory Infection including COVID-19 and adjusted for SARS-CoV-2 testing delay; lasso, least absolute shrinkage and selection operator; EMRR, Excess Mortality Rate Ratio of the number of SARI deaths in 2020, corrected for SARS-CoV-2 testing delay, to the number expected by Poisson regression; LB, Lower bound of the 95% confidence interval; UB, Upper bound of the 95% confidence interval.

period (0.012). For the mortality trend due to the natural causes, the coefficients were 0.09 for neoplasms, 0.03 for endocrine, nutritional/metabolic diseases and −0.09 for certain conditions originating in the perinatal period.

In summary, across different data sources and statistical methods, the following COVID-19 MR per 100 000 were calculated: 115 by the direct method adjusted for the testing delay, 91–96 by five regression methods estimating the excess SARI deaths in 2020, 104 by LNM with and 95 without lasso adjustment for excess mortality from natural causes.

## Discussion

To the best of the author's knowledge, this is the first paper on the under-reporting of COVID-19 deaths in Brazil and its federal states for the whole year of 2020 based on comparison of various methods and data sources. Other papers on this topic used the data as of June [20, 21], July [22], October [23] and September [24]. Except for the latter, all other used historical time series forecast [19], or 3-year average prior to the epidemic [21], or the last year before the epidemic [22] as a reference to estimate the EM due to COVID-19. Only two publications provided results for all federal states, as well as for the whole country [23, 25] but did not compare the results by multiple methods and data sources.

The indirect method results based on SARI were 17–21% lower than the MR of 115 per 100 000 obtained by the direct method. Excess mortality from the natural causes is a less specific indicator of COVID-19 deaths

**Table 4** Observed versus expected mortality from natural causes in Brazil, 2020

| | Not lasso-adjusted | | | Lasso-adjusted LNM† | |
|---|---|---|---|---|---|
| Year | Observed | Expected | Diff (%) | Expected | Diff (%) |
| 2009 | 964 391 | 970 909 | −0.67 | 976 509 | −1.24 |
| 2010 | 993 691 | 991 920 | 0.18 | 996 222 | −0.25 |
| 2011 | 1 024 656 | 1 013 655 | 1.09 | 1 018 106 | 0.64 |
| 2012 | 1 029 153 | 1 035 859 | −0.65 | 1 038 638 | −0.91 |
| 2013 | 1 058 791 | 1 059 161 | −0.03 | 1 060 328 | −0.14 |
| 2014 | 1 070 097 | 1 082 370 | −1.13 | 1 082 525 | −1.15 |
| 2015 | 1 112 039 | 1 106 078 | 0.54 | 1 104 496 | 0.68 |
| 2016 | 1 153 913 | 1 130 299 | 2.09 | 1 127 856 | 2.31 |
| 2017 | 1 154 006 | 1 155 040 | −0.09 | 1 151 103 | 0.25 |
| 2018 | 1 165 905 | 1 180 235 | −1.21 | 1 174 376 | −0.72 |
| 2019 | 1 205 432 | 1 206 079 | −0.05 | 1 198 384 | 0.59 |
| 2020 | 1 434 838 | 1 232 520 | 16.42 | 1 214 500 | 18.14 |

Diff, Difference observed vs. expected.

†Inferential lasso used for adjustment in linear regression with log-normal model and 10-fold cross-validation.

than the excess SARI mortality, and even more so regarding the direct method. The latter was double-checked for duplicated records and inconsistencies [5], thus considered the most accurate estimate of COVID-19 mortality and a benchmark against which other estimates are compared. The variation between these estimates across different data sources and statistical methods is not surprising given large uncertainties in diagnosing and reporting deaths from COVID-19 [26].

An earlier Brazilian study used excess mortality by natural causes up to mid-October 2020 and estimated the COVID-19 MR at 118 per 100 000 [23], not far from the present study results. Another study found the mortality by natural causes in Brazil 22% higher than expected as of early June 2020 compared to the 2015–2019 period [27], which is equivalent to the excess MR of 125 per 100 000. The most recent publication on this topic [25] estimated 57 070 undisclosed COVID-19 deaths in 2020, corresponding to about 23% downward bias and COVID-19 MR of 117 per 100 000 compared with the Ministry of Health data [5]. As COVID-19 treatment advanced, its case fatality reduced and brought about lower mortality by the end of the year, as suggested in the present study.

In the first months of the epidemic, excess all-cause mortality was suggested as a means to evaluate the impact of COVID-19 on mortality [28] and applied in some studies [29, 30]. In the Italian province of Lombardy, a 50% under-reporting of COVID-19 deaths was found [31], a value similar to that of five Brazilian state capitals with the highest incidence of COVID-19 [30]. On the other hand, the corresponding value for the USA was estimated at 26.3% [32]. However, all-cause mortality has at least two components: a direct influence of COVID-19 (e.g. respiratory failure) and an indirect influence (e.g. by delaying necessary treatment for other

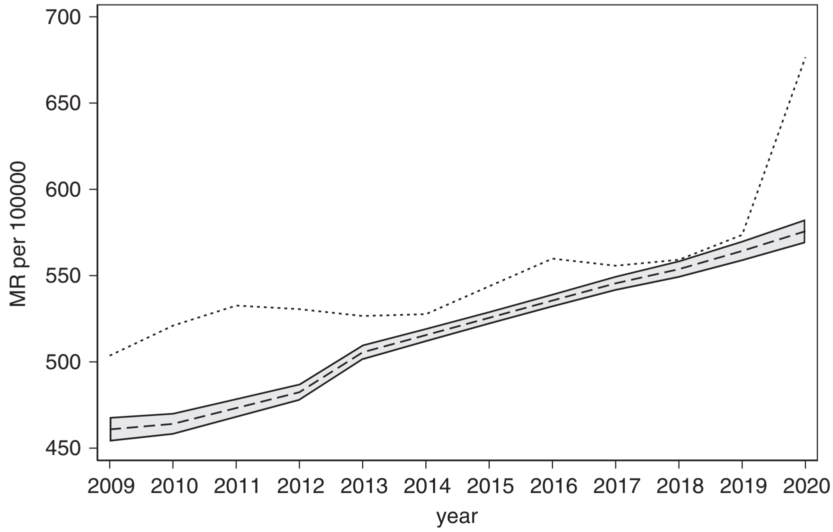

**Figure 1** Observed (dotted) and Poisson-expected (dashed line) mortality rate (MR) from natural causes in Brazil, 2020. Note: Shaded area represents 95% confidence interval for the MR predicted by lasso Poisson regression that accounted for confounding.

diseases). Therefore, excess all-cause mortality is the least specific indicator of the direct impact of COVID-19 on mortality and is applicable in the countries with timely and accurate notification of the causes of death.

About 3 of 4 SARI deaths were caused by COVID-19, with respective MR per 100 000 of 115 and 152 (Table 3). On the country level, both unadjusted and lasso-adjusted estimates pointed to a similar contribution of COVID-19 to increased SARI mortality in 2020, but less so on the state level.

The present study finding of 21.62% under-reporting of COVID-19 deaths in Brazil in 2020 by the direct method is close to the 19.7% found for the USA by mid-November 2020 [33]. However, an Indian study estimated a six times larger number of COVID-19 deaths than official reports [34], which amounts to almost 86% under-reporting. Many studies on the under-reporting of COVID-19 focused on the number of cases [35–38] as these are important for predicting future transmission rates. However, mortality remains the key parameter from the burden-of-disease perspective.

The 21.62% underestimation of COVID-19 mortality is very close to the 22% [23] and not far from 28% [25] excess mortality from natural causes and SARI, respectively. Two other studies found an average under-reporting of 40.68% (range 25.9–62.7%) for six large metropolitan areas in Brazil [20], and 30-57% in the state of Minas Gerais by mid-June 2020 compared with 2017–2019 mean [22]. The lower bound of these studies was close to the present study mean under-reporting estimate. The states with higher mortality rates are geographically scattered over the northern, northeastern and southeastern regions of Brazil, with Rio de Janeiro leading the ranking, in line with the findings of a nationwide SARS-CoV-2 antibody survey [39].

On the technical side, the SC method requires at least a moderate sample size relative to the number of variables in the model and may be vulnerable to overfitting regression models with many control cohorts [40]. Lasso regression is more flexible in fulfiling these requirements as it can fit a large number of variables, including polynomials and interactions of the control cohorts, thus achieving large predictive power regarding the outcome even with sparse data such as annual counts per state, whereas principal component analysis proposed to reduce sparsity for synthetic control method [10] is still limited to a linear combination of cohort variables. Lasso is less prone to overfitting, consistent and has good finite-sample properties, especially when combined with cross-validation [13, 14].

Although COVID-19 affected virtually all aspects of health care on a global scale and thus made it extremely difficult to apply instrumental variables and/or SC method in pre–post epidemic trend analysis [11], the present study used methods that do not depend on trends and covariates before COVID-19 epidemic, in addition to those that do, to evaluate the COVID-19 mortality underestimation. Unadjusted SARI trend analysis (Tables 2 and 3) with various statistical methods all pointed out the number of deaths significantly above the level before the epidemic. A difference of <8% (2.25 vs. 2.08) was found between the lasso-adjusted and unadjusted excess SARI MR ratio (Table 3). The key finding of 22.62% underestimation and corrected COVID-19 MR in Brazil in 2020 was based on the direct method, thus independent of the pre-epidemic data and eventual bias in the adjustment methods. Finally, a reasonable agreement between this result and that from the other two studies with the same scope [23, 25] provides some reassurance as to the validity of the conclusion.

Several limitations of the present study should be kept in mind. First, an important repository of respiratory viral infection data in Brazil was not included in the analysis because of significant delay in receiving SARI/COVID-19 notifications [3], despite a mathematical adjustment developed before the COVID-19 epidemic [41]. Second, the bias reporting COVID-19 deaths analysed here does not account for misdiagnosis of the causes of death, false-negative SARS-CoV-2 test results, or the unavailability of such tests [26], so that true downward bias is certainly larger. To illustrate the magnitude of misdiagnosis, it is worth noting an in-depth investigation of respiratory failure as causes of death nationwide in 2017 that found only 46.2% of these should be maintained as such [42]. Likely, intervening and intermediate causes of death are often reported where COVID-19 should be stated as the underlying cause [26]. Third, the imprecision of reported data was underestimated by the confidence intervals used but could be more adequately expressed with sensitivity analysis to be added in future research. For example, in 2017 the under-notification of death certificates based on civil registries in Brazil varied between 27.9% in the state of Maranhão to 0.5% in the Federal District, whereas the range reduced to 5.3% in Amapá to 0.3% in the São Paulo state when the Ministry of Health data were verified [43] [IBGE technical note]. No attempt was made to explain the reasons for state-wise variation in the present study as it was beyond its scope. The same goes for the lasso coefficients whose direct substantial interpretation is not supported due to their machine learning nature.

In 2019, the all-cause MR in Brazil was 642 per 100 000 inhabitants (1 348 232/210 147 125) [44] and could be a reasonable estimate for the year 2020 without

COVID-19 after correcting for the 1% annual increase in the last decade [43], thus resulting in the expected MR of 648 on the same scale. According to the direct method that focused solely on the direct impact of COVID-19 on mortality, the latter contributed almost 18% (115/648 = 0.177) of all-cause mortality, second only to cerebrovascular and ischaemic heart diseases. However, when the effect of COVID-19 included its total impact on the deaths from the natural causes (e.g. by aggravating pre-existing co-morbidities), the contribution of this disease reached a stunning 28–29%. It is therefore imperative that already available anti-COVID-19 vaccines are applied without delay.

## Conclusion

In Brazil, under-reporting of SARI, and especially SARS-CoV-2, is due to a low laboratory testing rate, reporting delay, inadequate access to medical care, and its poor quality, leading to the low sensitivity of epidemiological surveillance and poor outcomes, often without laboratory confirmation of the cause of infection and death. Based on the comparison of various statistical methods (exponential moving average, log-normal and Poisson regression with and without lasso adjustment), outcomes (COVID-19 alone, SARI, the natural causes of death) and data sources (Ministry of Health, nationwide death registries, state health authorities' on-line data), the best yet a still conservative estimate of under-reporting of COVID-19 deaths in 2020 was 22.62%. After correcting for this bias, the corresponding MR per 100 000 was 115 by the direct method and somewhat lower by two indirect methods based on the excess mortality of SARI and the natural causes in 2020. COVID-19 was the third cause of mortality that contributed directly to almost 18%, and indirectly with an additional 10-11%, to the death total in Brazil in 2020.

## References

1. Ministério da Saúde (Brazilian Ministry of Health). Painel Coronavírus. Ministério da Saúde (Available from: https://covid.saude.gov.br/) [12 January 2021].
2. WHO. Coronavirus Disease (COVID-19) Dashboard. WHO (Available from: https://covid19.who.int/) [12 January 2021].
3. Ministério da Saúde (Brazilian Ministry of Health). Sistema de Informação de Vigilância Epidemiológica da Gripe, Secretaria de Vigilância em Saúde. Monitoramento de casos de Síndrome Respiratória Águda Grave (SRAG) notificados no SIVEP-Gripe. (Available from: http://info.gripe.fiocruz.br) [15 June 2020].
4. Portal de Transparência. Registro Civíl. *Especial COVID-19*. Portal de Transparência. (Available from: https://transparencia.registrocivil.org.br/dados-covid-download) [21 January 2021].
5. Ministério da Saúde (Brazilian Ministry of Health). Boletins epidemiológicos. (Available from: https://coronavirus.saude.gov.br/boletins-epidemiologicos) [6 January 2021].
6. Pearl J. *Causality: Models, Reasoning & Inference*, 2nd Edition. Cambridge University Press: Cambridge, 2009.
7. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J Am Stat Assoc* 2010: **105**: 493–505.
8. Bonander C. Compared with what? Estimating the effects of injury prevention policies using the synthetic control method. *Inj Prev* 2018: **24**: i60–i66.
9. Bruhn CA, Hetterich S, Schuck-Paim C *et al*. Estimating the population-level impact of vaccines using synthetic controls. *Proc Natl Acad Sci USA* 2017: **114**: 1524–1529.
10. Shioda K, Schuck-Paim C, Taylor RJ *et al*. Challenges in estimating the impact of vaccination with sparse data. *Epidemiology* 2019: **30**: 61–68.
11. Shioda K, Weinberger DM, Mori M. Navigating through health care data disrupted by the COVID-19 pandemic. *JAMA Intern Med* 2020: **12**: 1569–1570.
12. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B* 1996: **1996**: 267–288.
13. Belloni A, Chernozhukov V, Hansen C. High-dimensional methods and inference on structural and treatment effects. *J Econ Perspect* 2014: **28**: 29–50.
14. Chernozhukov V, Chetverikov D, Demirer M *et al*. Double/debiased machine learning for treatment and structural parameters. *Econom J* 2018: **21**: C1–C68.
15. Becketti S. *Introduction to Time Series Using Stata*. Stata University Press: College Station, TX, 2010.
16. Center for Disease Control and Prevention (CDC). Comprehensive Listing ICD-10-CM Files. January 1, 2021 release of ICD-10-CM. CDC (Available from: https://www.cdc.gov/nchs/icd/icd10cm.htm) [12 January 2021].
17. Ministério da Saúde (Brazilian Ministry of Health). OpenDataSUS. Dados de Corona Virus (Available from: https://opendatasus.saude.gov.br/group/dados-do-coronavirus) [January 15, 2021].
18. Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics). Projeção populacional (Population projection). (Available from: https://www.ibge.gov.br/apps/populacao/projecao/index.html) 2021.
19. StataCorp. *Stata: Release 16. Statistical Software*. StataCorp LLC: College Station, TX, 2019.
20. Veiga e Silva L, de Andrade Abi Harb MDP, Teixeira Barbosa dos Santos AM *et al*. COVID-19 mortality underreporting in Brazil: analysis of data from government internet portals. *J Med Internet Res* 2020: **18**: e21413.
21. Alves THE, de Souza TA, Silva SA *et al*. Underreporting of death by COVID-19 in Brazil's second most populous state. *Front Public Health* 2020: **8**: 578–645.

22. Amaral PHR, Andrade LM, da Fonseca FG *et al*. Impact of COVID-19 in Minas Gerais, Brazil: excess deaths, sub-notified cases, geographic and ethnic distribution. *Transbound Emerg Dis* 2020. https://doi.org/10.1111/tbed.13922.
23. Carvalho TA, Boschiero MN, Marson FAL. COVID-19 in Brazil: 150,000 deaths and the Brazilian underreporting. *Diagn Microbiol Infect Dis* 2021: **99**: 115258.
24. Silva L, Figueiredo FD. Using Benford's law to assess the quality of COVID-19 register data in Brazil. *J Public Health (Oxf)* 2021: **43**: 107.
25. Paixão B, Baroni L, Pedroso M *et al*. Estimation of COVID-19 under-reporting in the Brazilian states through SARI. *New Gener Comput* 2021: **14**: 1–23. https://doi.org/10.1007/s00354-021-00125-3.
26. França EB, Ishitani LH, Teixeira RA *et al*. Deaths due to COVID-19 in Brazil: how many are there and which are being identified? *Rev Bras Epidemiol* 2020: **22**(Suppl 3): e190010.
27. Marinho F, Torrens A, Teixeira R *et al*. Aumento das mortes no Brasil, Regiões, Estados e Capitais em tempo de COVID-19: excesso de óbitos por causas naturais que não deveria ter acontecido. (Available from: https://www.-google.com/search?client=firefox-b-d&q=Aumento+das+mortes+no+Brasil%2C+Regi%C3%B5es%2C+Estados+e+Capitais+...www.vitalstrategies.org+%E2%80%BA+wp-content+%E2%80%BA+uploads) [12 January 2021].
28. WHO: Revealing the Toll of COVID-19: A Technical Package for Rapid Mortality Surveillance and Epidemic Response. WHO (Available from: https://www.who.int/publications/i/item/revealing-the-toll-of-covid-19) [21 May 2020].
29. Mannucci E, Nreu B, Monami M. Factors associated with increased all-cause mortality during the COVID-19 pandemic in Italy. *Int J Infect Dis* 2020: **98**: 121–124.
30. Freitas ARR, Medeiros NM, Frutuoso LCV *et al*. Tracking excess deaths associated with the COVID-19 epidemic as an epidemiological surveillance strategy-preliminary results of the evaluation of six Brazilian capitals. *Rev Soc Bras Med Trop* 2020: **53**: e20200558.
31. Buonanno P, Galletta S, Puca M. Estimating the severity of COVID-19: Evidence from the Italian epicenter. *PLoS One* 2020: **15**: e0239569.
32. Stokes AC, Lundberg DJ, Hempstead K *et al*. Assessing the impact of the covid-19 pandemic on us mortality: a county-level analysis. medRxiv Preprint 2020: 2020.08.31.20184036.
33. Angulo FJ, Finelli L, Swerdlow DL. Estimation of US SARS-CoV-2 infections, symptomatic infections, hospitalizations, and deaths using seroprevalence surveys. *JAMA Netw Open* 2021: **4**: e2033706.
34. Bhaduri R, Kundu R, Purkayastha S *et al*. Extending the Susceptible-Exposed-Infected-Rremoved (SEIR) Model to handle the high false negative rate and syptom-based administration of COVID-19 diagnostic tests: SEIR-fansy. *medRxiv* Preprint 202: 2020.09.24.20200238.
35. Prado MFD, Antunes BBP, Bastos LDSL *et al*. Analysis of COVID-19 under-reporting in Brazil. *Rev Bras Ter Intensiva* 2020: **32**: 224–228.
36. Russell TW, Golding N, Hellewell J *et al*. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Med* 2020: **18**: 332.
37. de Oliveira ACS, Morita LHM, da Silva EB *et al*. Bayesian modeling of COVID-19 cases with a correction to account for under-reported cases. *Infect Dis Model* 2020: **5**: 699–713.
38. Unnikrishnan J, Mangalathu S, Kutty RV. Estimating under-reporting of COVID-19 cases in Indian states: an approach using a delay-adjusted case fatality ratio. *BMJ Open* 2021: **11**: e042584.
39. Hallal PC, Hartwig FP, Horta BL *et al*. SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. *Lancet Glob Health* 2020: **8**: e1390–e1398.
40. Potscher BM, Leeb H. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J Multivar Anal* 2009: **100**: 2065–2082.
41. Bastos LS, Economou T, Gomes MFC *et al*. A modelling approach for correcting reporting delays in disease surveillance data. *Stat Med* 2019: **38**: 4363–4377.
42. França EB, Ishitani LH, Teixeira RA *et al*. Improving the usefulness of mortality data: reclassification of ill-defined causes based on medical records and home interviews in Brazil. *Rev Bras Epidemiol* 2019: **23**: e200053.
43. Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics) *Nota Técnica* (Technical Note) (Available from: https://www.ibge.gov.br/estatisticas/sociais/populacao/26176-estimativa-do-sub-registro.html?edicao=26182&t=o-que-e) [5 July 2020].
44. Ministério da Saúde (Brazilian Ministry of Health). Portal da Saúde SUS. Informações de Saúde (TABNET) (Available from: http://www2.datasus.gov.br/DATASUS/index.php?area=02) [12 January 2021].

**Corresponding Author** **Emil Kupek**, Federal University of Santa Catarina, CCS, Department of Public Health, 88040-900 Florianopolis/SC, Brazil. E-mail: emilkupek@gmail.com