Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil

William Marciel de Souza [1]*, Lewis Fletcher Buss [2]*, Darlan da Silva Candido [3]*, Jean-Paul Carrera [3,4]*, Sabrina Li [5]*, Alexander E. Zarebski [3], Rafael Henrique Moraes Pereira [6], Carlos A. Prete Jr [7], Andreza Aruska de Souza-Santos [8], Kris V. Parag [9], Maria Carolina T.D. Belotti[7], Maria F. Vincenti-Gonzalez [10], Janey Messina [5,11], Flavia Cristina da Silva Sales [2], Pamela dos Santos Andrade [2], Vítor Heloiz Nascimento [7], Fabio Ghilardi [2], Leandro Abade [3], Bernardo Gutierrez [3,12], Moritz U. G. Kraemer [3,13,14], Carlos K.V. Braga [6], Renato Santana Aguiar [15], Neal Alexander [16], Philippe Mayaud [17], Oliver J. Brady [15,18], Izabel Marcilio [19], Nelson Gouveia [20], Guangdi Li [21], Adriana Tami [10], Silvano Barbosa de Oliveira [22], Victor Bertollo Gomes Porto [22], Fabiana Ganem [22], Walquiria Aparecida Ferreira de Almeida [22], Francieli Fontana Sutile Tardetti Fantinato [22], Eduardo Marques Macário [23], Wanderson Kleber de Oliveira [23], Mauricio L. Nogueira [24], Oliver G. Pybus [3], Chieh-Hsi Wu [25]*, Julio Croda [23,26,27,28]*#, Ester C. Sabino [2]*, Nuno Rodrigues Faria [3,29]*#

* These authors contributed equally

Corresponding author

1. Centro de Pesquisa em Virologia, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brazil.

2. Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.

3. Department of Zoology, University of Oxford, Oxford, UK.

4. Department of Research in Virology and Biotechnology, Gorgas Memorial Institute of Health Studies, Panama City, Panama.

5. School of Geography and the Environment, University of Oxford, Oxford, UK.
6. Institute for Applied Economic Research (IPEA), Brasília, Brazil.
7. Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil.
8. Brazilian Studies Programme, Latin American Centre, University of Oxford, Oxford, UK.
9. MRC Centre for Global Infectious Disease Analysis, Imperial College London, UK.
10. University of Groningen, University Medical Center Groningen, Department of Medical Microbiology and Infection Prevention, Groningen, Netherlands.
11. Oxford School of Global and Area Studies, University of Oxford, Oxford, UK.
12. School of Biological and Environmental Sciences, Universidad San Francisco de Quito USFQ, Quito, Ecuador.
13. Harvard Medical School, Harvard University, Boston, MA, USA.
14. Boston Children's Hospital, Boston, MA, USA.
15. Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
16. MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK.
17. Department of Clinical Research, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK.
18. Centre for the Mathematical Modelling of Infectious Diseases, Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK.
19. Núcleo de Vigilância Epidemiológica do Hospital das Clínicas da Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil.
20. Departamento Medicina Preventiva, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.

21. Department of Epidemiology and Health Statistics, Xiangya School of Public Health, Central South University, Changsha 410078, China.

22. Secretariat of Health Surveillance, Department of Immunization and Communicable Diseases, Brazilian Ministry of Health, Brasília, Brazil.

23. Secretariat of Health Surveillance, Brazilian Ministry of Health, Brasília, Brazil.

24. Faculdade de Medicina de São José do Rio Preto, São Jose do Rio Preto, Brazil.

25. Mathematical Sciences, University of Southampton, Southampton, UK.

26. Laboratório de Pesquisa em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados, Brazil.

27. Fundação Oswaldo Cruz, Campo Grande, Mato Grosso do Sul, Campo Grande, Brazil.

28. Department of Epidemiology of Microbial Diseases, Yale University School of Public Health, New Haven, USA.

29. Department of Infectious Disease Epidemiology, Imperial College London, London, UK.

Corresponding authors:

Nuno Rodrigues Faria, PhD

Email: nfaria@ic.ac.uk

Julio Croda, MD PhD

Email: juliocroda@gmail.com

Abstract

The first case of COVID-19 was detected in Brazil on February 25, 2020. We report and contextualize epidemiological, demographic, and clinical findings for COVID-19 cases during the first three months of the epidemic. By May 31, 2020, 514,200 COVID-19 cases, including 29,314 deaths had been reported in 75.3% (4,196 of 5,570) of municipalities across all five administrative regions of Brazil. R_0 for Brazil was estimated at 3.1 (95% BCI 2.4–5.5), with a higher median but overlapping credible intervals compared to some other seriously affected countries. A positive association between higher per-capita income and COVID-19 diagnosis was identified. Further, the severe acute respiratory infection cases with unknown aetiology were associated with lower per capita income. Co-circulation of six respiratory viruses was detected but at very low levels. These findings provide a comprehensive description of the ongoing COVID-19 epidemic in Brazil and may help guide subsequent measures to control virus transmission.

Introduction

Coronavirus disease 2019 (COVID-19) is a severe acute respiratory infection that emerged in early December 2019 in Wuhan, China[1]. The outbreak was declared a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO) on January 30, 2020. COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), an enveloped, single-stranded positive-sense RNA virus that belongs to the *Betacoronavirus* genus, *Coronaviridae* family[2]. SARS-CoV-2 is closely related genetically to bat-derived SARS-like coronaviruses[3]. Human-to-human transmission occurs primarily via respiratory droplets and direct contact, similar to human influenza viruses, SARS-CoV and Middle East Respiratory Syndrome virus (MERS-CoV)[4]. The most commonly reported clinical symptoms are fever, dry cough, fatigue, dyspnoea, anosmia, ageusia, or some combination of these[1,4,5]. As of June 16, 2020, more than 7.9 million cases have been confirmed worlwide, resulting in 434,796 deaths[6].

Brazil declared COVID-19 as a national Public Health Emergency (PHE) on February 3, 2020[7]. After the development of a national emergency plan and the early establishment of molecular diagnostic facilities across Brazil's network of public health laboratories, the country reported its first confirmed COVID-19 case on February 25, 2020, in a traveller returning to São Paulo from northern Italy[8]. São Paulo is the largest city in South America and no other Brazilian city receives a greater proportion of international flights[9]. Currently, Brazil has one of the fastest-growing COVID-19 epidemics in the world, now accounting for 1,864,681 cases and 72,100 deaths, comprising over 55% of the total number of reported cases in Latin America and Caribbean (as of July 14, 2020)[6]. About 21% of Latin American and Caribbean populations are estimated to be at risk of severe COVID-19 illness[10]. The region has been experiencing large outbreaks, with growing epidemics in Brazil, Peru, Mexico, Chile, Colombia, Panama, and possibly Venezuela and Nicaragua, amidst growing

concerns on testing capacity for COVID-19[11-14]. Preparedness for laboratory surveillance of SARS-CoV-2 in Latin America is centred around a network of national reference influenza surveillance laboratories that is facing several challenges, including shortage of reagents and equipment[15].

Conscious of the challenges associated with surveillance since the beginning of the epidemic in Brazil, here we focus on two main objectives. First, we contextualize the Brazilian SARS-CoV-2 epidemic by comparing local transmission dynamics with those observed in selected other countries. Second, we use geospatial data related to confirmed COVID-19 cases and severe acute respiratory infection (SARI) cases with unknown aetiology to evaluate the relationship between socio-economic factors and COVID-19 distribution.

Results

Contextualizing COVID-19 data notification systems in Brazil

On January 22, 2020, more than one month before the first case in Brazil, the Brazilian Ministry of Health implemented the REDCap platform to notify prospective suspected, probable, and confirmed COVID-19 cases (see **Methods** for case definitions), as part of early response to the pandemic[16]. By March 27, 2020, the REDCap system was discontinued (**Fig. 1**). Since then, mild-COVID-19 cases started to be notified on e-SUS-VE (e-SUS Vigilância Epidemiológica), a new national COVID-19 notification system and hospitalised COVID-19 cases started to be recorded on a pre-existing SIVEP-Gripe system. The SIVEP-Gripe system has been in use since 2009 (influenza H1N1 2009 pandemic) and has since centralized the notification of respiratory viruses and SARI for the Brazilian Ministry of Health (**Fig. 1**). Both the e-SUS-VE and SIVEP-Gripe include suspected and confirmed COVID-19 cases by public health and private services (primary and emergency care). These two notification systems (e-SUS-VE and SIVEP-Gripe) are inter-related on the *Portal do COVID-19* website (https://covid.saude.gov.br/), which summarises daily the aggregated counts from both platforms.

SARS-CoV-2 notification in Brazil: international transmission to rapid internal dissemination

We analysed a total of 514,200 SARS-CoV-2 cases from the *Portal do COVID-19* website (SIVEP-Gripe, and e-SUS VE databases combined) that were confirmed by molecular diagnostic and clinical epidemiological criteria by May 31, 2020 (see Materials and Methods). Cases were reported in 75.3% (4,196 of 5,570) of municipalities across all five administrative regions of Brazil and included 206,555 (40.2%) recovered patients, and 29,314 fatal (17.5%) COVID-19 cases (**Fig. 2A**). We further analysed a total of 1,468 confirmed

cases from the REDCap system, including 342 imported cases with associated travel history information. After excluding cases involving with that travelled to multiple countries before entering Brazil (*n*=56) and that had an unknown country of origin (*n*=16). The self-reported countries of infection for cases acquired abroad until March 19, 2020 were USA (28.6%, *n*=76), Italy (24.4%, *n*=65), and the United Kingdom (10.5%, *n*=28) and Spain (8.3%, *n*=22) (**Extended Data Fig. 1**). The first reported case (SPBR1) was reported on February 25, 2020 in the municipality of São Paulo, the fourth most populous urban area worldwide. Following the first notifications of COVID-19 in Brazil's largest population centres, we find that SARS-CoV-2 subsequently spread to municipalities with smaller population sizes (**Fig. 2B**). Until May 31, 2020, most confirmed cases and deaths were reported in the states of São Paulo (109,698 cases and 7,615 deaths), Rio de Janeiro (53,388 cases and 5,344 deaths), Ceará (48,489 cases and 3,010 deaths) and Amazonas (41,378 cases and 2,052 deaths), which together account for 49.2% of all cases and 61.5% of deaths in Brazil (**Fig. 2c**).

Basic reproduction number (R_0) of SARS-CoV-2 in Brazil and comparison countries

To estimate the basic reproduction number (R_0) of SARS-CoV-2 in Brazil, daily confirmed cases in São Paulo, Rio de Janeiro, Ceará and Amazonas states were compiled from the Ministry of Health (for specification of the time-windows used in the analyses see **Extended Data Fig. 2**). For comparison, we compiled time series of confirmed cases in several European countries from the Johns Hopkins Coronavirus Resource Center (https://coronavirus.jhu.edu/, see also **Extended Data Fig. 3**). We found that São Paulo, Rio de Janeiro and Amazonas were characterized by similar R_0 values of 2.9 (95% Bayesian credible interval, BCI, 2.2–5.1), 2.9 (95% BCI 2.2–4.9) and 2.6 (95% BCI 2.0–4.5). However, for Ceará, estimated R_0 was considerably lower, 1.9 (95% BCI 1.5–3.0) (**Fig. 3**, **Extended Data Fig. 1**). This finding could be a result of the small window between the first

notified cases and the early implementation of non-pharmaceutical interventions (NPIs) in this state (**Supplementary Table 1**, **Extended Data Fig. 2**). On a national scale, the estimated R_0 for Brazil was slightly higher than that of the Brazilian states considered in this study, with a median of 3.1 (95% BCI 2.4–5.5), and also slightly higher than R_0 values estimated for other severely affected countries: Spain (2.6, 95% BCI 2.0–4.6), France (2.5, 95% BCI 1.9–4.4), United Kingdom (2.6, 95% BCI 2.0–5.1) and Italy (2.5, 95% BCI 2.0–4.4) (**Fig. 3**). While the incidence curves for European countries have consistently flattened and declined after the implementation of NPIs (suggesting R_0 has fallen below one), Brazil's daily incidence curve has continued to increase (**Fig. 2A and Extended Data Fig. 4**).

Severe acute respiratory infections (SARI) mostly reflect COVID-19 cases

In the early-phase of the COVID-19 epidemic in Brazil, we analysed the results for other respiratory pathogens tested in Brazil as part of the differential diagnosis by Central Public Health Laboratories and National Influenza Centres (Brazilian Ministry of Health) obtained from a REDcap platform[17] designed for COVID-19. The respiratory viruses most frequently identified between January 2020 and March 27, 2020, in patients with suspected but negative diagnosis of COVID-19 were influenza A virus (347 [14.3%] of 2,429 tested cases), influenza B virus (251 [10.3%] of 2,429) and human rhinovirus (136 [5.6%] of 2,429). We found co-detection of SARS-CoV-2 with six other respiratory viruses, the most frequently were influenza A (11 [0.5%] of 2,429) and human rhinovirus (6 [0.2%] of 2,429) (**Extended Fig. 7**).

The SIVEP-Gripe system started reporting hospitalised COVID-19 cases in early March 2020 (epidemiological week 10) (**Fig. 4**). In this system, the number of tested cases is unavailable. We found that the peak of influenza confirmed cases (n=447) occurred at

epidemiological week 12 (15-21 March 2020). During the same week 12, we detected an 8.5-fold increase in total cases attributed to SARS-CoV-2 (*n*=3,789) and a 9.9-fold increase in total cases notified as SARI with unknown aetiology (*n*=4,424) (**Fig. 4**). From January to May 31, 2020, a total of 2,136 influenza cases and 272 cases caused by other respiratory pathogens including human respiratory syncytial virus, human rhinovirus, adenovirus, metapneumovirus were notified in the SIVEP-Gripe database. The low observed incidence of influenza and other respiratory viruses may be influenced by limited testing for these viruses during this period. Although NPIs may have an impact in reducing influenza virus transmission, this does not necessarily reflect a lower co-circulation of other respiratory viruses[18].

Socio-economic differences are associated with COVID-19 diagnosis

Until 31 May 2020, a total of 73,648 COVID-19 confirmed cases and 168,001 SARI cases with unknown aetiology were notified in the SIVEP-Gripe system. We hypothesized that the 2.3-fold increase of SARI cases with unknown aetiology was associated with differential access to healthcare due to socio-economic factors.

We focus on the Metropolitan Region of São Paulo (MRSP) that has a population of 23 million inhabitants across 6 sub-regions (Central, West, North, East, Southeast and Southwest) and 39 municipalities (**Fig. 5A**). To test this hypothesis, we obtained *per capita* income at the census tract level (typically 150-300 households) in the MRSP, based on the residential address of each case. We then linked this information to each patient's final diagnosis outcome: COVID-19 confirmed case or SARI with unknown aetiology. While the income distribution of SARI cases with unknown aetiology was similar to that of the MRSP over the whole period (**Fig. 5B**), we observed that the income distribution individuals conformed to be COVID-19-cases confirmed by laboratory and clinical criteria was initially

higher and decreased over time towards the distribution for the whole of the MRSP by epidemiological week 21 (**Fig. 5B**). Importantly, we found that the log odds of one or more confirmed COVID-19 case per census tract increased with per capita income in epidemiological weeks 12 and 22 (likelihood ratio test [LRT] *P*-value <0.001 (**Fig. 5B** and **Supplementary Table 2**). This provides statistical evidence of an association between confirmed COVID-19 diagnosis and *per capita* income, suggesting a socio-economic difference in access to COVID-19 diagnosis in the MRSP. For reference, we also provide a map of per capita income (**Fig. 5A**) and population density in each census tract (**Extended Data Fig. 8**).

We conducted a geospatial analysis to understand the distribution of relative risk of observing a COVID-19 case or an SARI cases with unknown aetiology in the MRSP, using a Bayesian method and adjusted for spatial and non-spatial effects defined by Besag-York-Mollié model[19] (**Fig. 5**). Our estimates show an increase in the relative risk of COVID-19 diagnosis in higher income census tracts between epidemiological weeks 12 to 21, especially in the central region of the MRSP (**Figs. 5A and 5C**). We observed a similar trend in the relative risk of SARI cases with unknown aetiology among residents of the central region. However, there is also increased probability of SARI cases with unknown aetiology in the southwest, west, north, and south sub-regions, where income per capita is typically lower. Overall, the relative risk of SARI cases with unknown aetiology is more spatially widespread in the MRSP than of confirmed COVID-19 cases (**Fig. 5C**).

The relative risk of SARI cases with unknown aetiology compared to confirmed COVID-19 cases in the central region of the MRSP decreases through time likely as a response to several NPIs implemented throughout the state of São Paulo (see **Supplementary Table 1**). By week 16, one month after the start of the NPIs in São Paulo, we detected an increased risk particularly of SARI cases with unknown aetiology outside the central region

of the MRSP, especially in the southwest region. SARI cases with unknown aetiology risk was also high in the east region. By week 21, the risk remained high throughout the central region and SARI cases with unknown aetiology risk decreased in the east region, possibly as a result of interventions targeting the reduction of SARS-CoV-2 transmission.

Demographics and characteristics of COVID-19 hospitalised and fatal cases in Brazil

Analysis of the age-sex structure of 67,180 confirmed COVID-19 cases notified on the SIVEP-Gripe system revealed a high proportion (44,027 [65.5%] of 67,180) of confirmed COVID-19 infections in middle or older-age individuals (≥50 years of age) and a lower proportion (1,454 [2.2%] of 67,180) in younger age groups (≤ 20 years of age) (**Fig. 6A**). The median age was 59 years (IQR = 44–72). The majority (38,654 [57.5%] of 67,180) were male. Similarly, 59% (14,498 of 24,519) of COVID-19 deaths were in men, and 85% (20,916 of 24,519) were in people aged ≥50 years. A total of 2.95% (1,983 of 67,180) cases were reported as nosocomial transmission, defined as a COVID-19 case acquired after hospitalization. Overall, 116 newborns (≤ one month old), 381 infants (≥1 to 12 month-old), 518 children (≥1 to 12 years old), and 258 adolescents (≥12 to 17 years of age) were diagnosed with COVID-19. In addition, 740 patients were pregnant, 61 in the first trimester, 172 in the second trimester, 447 in the third trimester, and 60 had missing gestational age.

By 31 May 2020, 91% (67,042 of 73,649) of patients with COVID-19 notified in the SIVEP-Gripe system had been hospitalized. Of these, 30.3% (22,332 of 73,649) were admitted to an intensive care unit (ICU). The median length of ICU stay for COVID-19 patients was five days (IQR, 2–10, range: 0-65 days), based on the ICU admission and discharge dates of 8,240 confirmed cases. Most symptoms reported by COVID-19 patients were cough (56,681 [85.2%] of 66,514 without missing data), fever (51,312 [79.6%] of

65,310) and dyspnoea (51,312 [76.6%] of 65,310) (**Fig. 6B**). These three symptoms compose part of the case definition of SARI in Brazil. In addition, 68% (40,806 of 60,400) of COVID-19 cases were hypoxic (O_2 saturation < 95%) reflecting the overall severity of cases notified on SIVEP-Gripe (as shown in **Fig. 1)**. The most prevalent comorbidities were cardiovascular disease (23,085 [66.5%] of 34,693 without missing data) and diabetes (17,271 [54.5%] of 31,672) (**Fig. 6A**). Among the COVID-19 patients, older age groups tended to have a higher proportion of comorbidities than younger age groups in different outcomes (**Fig. 6C**). The proportions of the general Brazilian population with cardiovascular disease and diabetes are 4.2%, and 6.2%, respectively[20]. A total of 83.7% (17,921 of 21,414 with complete comorbidity information) confirmed COVID-19 cases had at least one comorbidity (see **Supplementary Table 2** for information on data completeness).

Discussion

While the COVID-19 epidemic in Brazil continues to grow, details of its transmission potential and clinical and epidemiological characteristics remains poorly understood. We estimate a higher median transmission potential, R_0 of 3.1 (2.4–5.5), of SARS-CoV-2 in Brazil compared with Italy, UK, France, and Spain, which have point estimates of R_0 varying from 2.5 to 2.6, however the credible intervals overlap substancially. We have also observed rapid spread of COVID-19 through the country, with more populated and better-connected municipalities being affected earlier and less populated municipalities being affected at a later stage of the epidemic. In the São Paulo metropolitan region, we found a higher risk of diagnosed COVID-19 cases in census tracts with higher per capita income during the early-phase of COVID-19 epidemic but also as weeks progressed. This contrasts with the wider spread of SARI cases among sub-regions with lower per capita income. Our results provide new insights into the Brazilian COVID-19 epidemic and highlight the high transmission potential of SARS-CoV-2 in the country, the role of its large urban centres, and the lack of lockdown, the challenges in notification and non-equitable access to testing/diagnostic as factors potentially contributing to the rapid and sustained spread of the epidemic in Brazil.

Recent estimates of R_0 at the beginning of the COVID-19 epidemic in Brazil have suggested that an infected individual would infect on average three or four others[21]. The credible intervals of our estimates broadly overlap with these observations and are lower compared to previously published estimates for Brazil[22]. As a comparison, reproduction number in Peru have been estimated at around 2.3 (2.0–2.5)[23]. Since the start of the epidemic in Brazil, several types of NPI have been adopted with varied success by the country's 27 federal units and 5,596 municipalities. Virus transmission seems to have dropped substantially in most affected states[21] and also in the city of São Paulo[24]. However, the estimated reproduction number remains above one[21,24]. Thus, only mitigation (and not

304 suppression) of the epidemic has been achieved so far, which has been linked to substantial
305 excess deaths due to poorer health care available[25,26]. Closer surveillance of viral
306 transmission at the local scales and an assessment of the impact of the different control
307 measures on COVID-19 transmission will help to determine a "optimal" mitigation strategy
308 to minimize infections and reduce healthcare demand in Brazil. Moreover, continued
309 monitoring of the genetic diversity of the virus lineages circulating in Brazil[24] will be
310 important, as recent data suggests that virus diversity may play a role in virus
311 transmissibility[27,28].

312 We find that 65.5% of notifications in the SIVEP-Gripe system, which includes most
313 severe COVID-19 cases are from patients aged ≥50 years of age. This observation is
314 remarkably similar to current estimates for Latin America[10], where 65% of the individuals
315 ≥50 years of age have been estimated to be at high risk of severe COVID-19, defined as
316 individuals with at least one condition who would require hospitalisation if infected.
317 Moreover, we find that 57% and 59% of the severe COVID-19 cases and deaths
318 (respectively) notified in SIVEP-Gripe were male, and that the most frequent comorbidities
319 were cardiovascular disease and diabetes. Overall 84% of SIVEP-Gripe notifications had at
320 least one underlying condition; of these, 21% (*n*=9,471/45,480) are included in the working
321 age (16 to 65 years of age). Moreover, only 2.6% (n=1892/73,673) of the COVID-19
322 confirmed cases notified in the SIVEP-Gripe system include occupation. Information on
323 socio-economic determinants as well as occupation and race/ethnicity are critical[29] as this
324 allows to prioritisation of control efforts, for example towards healthcare workers and
325 patients attending hospitals[30] or work settings[31].

326 Our data uncovers a socio-economic bias in testing and diagnostics in current
327 surveillance guidelines and suggests that the number of notified confirmed case counts may

328 substantially underestimate the number of cases in the general population, particularly in
329 regions of lower socio-economic status. Socio-economic differences are associated with
330 access to healthcare[32] and should be taken into account when designing targeted
331 interventions. We find that the proportion of SARI cases with unknown aetiology to
332 confirmed COVID-19 cases has increased across the entire country (as of June 15, 2020, the
333 number of notified SARI cases with unknown aetiology is nearly 2-fold greater than
334 confirmed COVID-19 cases). Based on clinical and epidemiological grounds, it is likely that
335 many SARI cases with unknown aetiology are caused by SARS-CoV-2. In order rigorously
336 establish the contribution of non-SARS-CoV-2 infections to the SARI cases, we would need
337 additional denominator data to understand the level of testing for these viruses, i.e., the
338 negative test results. Our findings with regards to socio-economic bias are likely to apply to
339 other states and regions of Brazil and highlight the importance of scaling up surveillance and
340 laboratory capacity within Latin America. Indeed, the largest Brazilian serosurvey conducted
341 to date suggests that undetedected cases may be seven times higher than reported cases[33].

342 We further show that SARI cases with unknown aetiology are associated with lower
343 socio-economic status in the Metropolitan Region of São Paulo. The socio-economic
344 disparities observed here were particularly evident at the beginning of the outbreak (**Fig. 5B**).
345 This can be explained in part by (*i*) the high proportion of early cases in returning travellers
346 with higher income and better access to private laboratories for diagnostics, and (*ii*) the more
347 limited access to freely available diagnostic screening. For example, between February 25
348 and March 18, 2020, two thirds (586 [66.9%] of 876) of diagnostic tests were performed in
349 private medical laboratories where costs varied typically between 300-690 Brazilian Reais
350 (BRL) (for context, current minimum monthly salary is 1,045 BRL). Thus, the true burden of
351 the epidemic in lower income neighbourhoods is most likely underestimated. In New York
352 City, for example, poorer neighbourhoods had higher disease burden, driven in part by the

353 movement of essential workers using public transport during the pandemic[34]. Data-driven
354 analyses are urgently needed to help tackling health inequities during the ongoing epidemic
355 in Brazil. Strategies to evaluate and control transmission should consider differential assess to
356 COVID-19 diagnosis for lower income populations, changes in notification systems and
357 delays in reporting which are key to accurately determine rates of epidemic growth[35].
358 Innovative infectious disease surveillance approaches such as those obtained from aggregated
359 mobility data, when used properly, could help supporting public health actions across the
360 COVIV-19 epidemic[36-39].

361 Epidemics of COVID-19 and influenza seem to have occurred simultaneously in
362 Brazil (**Fig.4** and **Extended Data Figure 7**) and symptoms overlap between the two
363 infections. We detected co-circulation of eight other respiratory viruses, the most common
364 respiratory infections were influenza A and B, and human rhinovirus. We also detected
365 multiple co-detection of SARS-CoV-2 with other respiratory viruses, such as influenza A, B
366 and human metapneumovirus, which have also been reported elsewhere[40,41]. Although, co-
367 infections with other respiratory viruses have been reported in other countries[42-44], no
368 difference in clinical disease severity between cases with and without viral co-infection has
369 been observed thus far[45]. The co-circulation of other respiratory pathogens highlights the
370 need of scaling up laboratory and molecular screening of SARS-CoV-2 and other respiratory
371 viruses in public laboratories across Brazil[15]. Continued molecular and genomic surveillance
372 will be important to determine patterns of virus transmission and guide public health
373 measures in forthcoming phases of the epidemic[24,46-48].

374 There are several limitations to this study. First, detailed individual-level data were
375 only available for REDcap and SIVEP-Gripe systems, in which many cases had incomplete
376 documentation, particularly regarding comorbidities. Second, our socio-economic analysis
377 was based partially on ecological inference, using the *per capita* income in the census tract of

residence (tather than the actual income of the patients), and assuming the same denominator for each census tract (~300 households). We emphasize that our spatial analysis is prone to metholodological constraints caused by ecological fallacy and the modifiable areal unit problem. These constraints are inherent to any spatial analysis of aggregated data. Despite the above-mentioned limitation, census tract corresponds to small areas of analysis, of no more than 300 households but often less than that. Social science literature on Brazil not only highlights the country's socio-economic inequality but also how it is spatially pronounced, for that reason, census tract remains a useful tool to infer *per capita* income in the absence of individual-level data. In addition, our databases were predominantly composed of hospitalised COVID-19 patients, and we were unable to evaluate the rate of hospitalisation among the different socio-economic status. In the future, robust modelling of the relationships between socio-economic factors and disease severity will require a data collection system with detailed information on symptoms/signs and comorbidities both in severe and non-severe cases. Finally, our retrospective study has focused predominantly on symptomatic patients that presented or were referred to health services for testing. Therefore, we are unable and do not attempt to describe the full spectrum of disease, nor can we describe the full epidemiological picture of this epidemic.

In conclusion, we have provided a comprehensive assessment of COVID-19 notification and transmission in Brazil. Our findings provide important context for diagnostic screening and health-care planning, and for future precision studies focussing on the impact of non-pharmaceutical and pharmaceutical interventions, and the effect of social health determinats on COVID-19 transmission.

Methods

Ethical approval and case definitions

This retrospective national study was supported by the Brazilian Ministry of Health and ethical approval was provided by the national ethical review board (Comissão Nacional de Ética em Pesquisa, CONEP), protocol number CAAE 30127020.0.0000.0068.

A patient presenting with an acute respiratory syndrome (fever and at least one sign/symptom of respiratory illness), and (i) history of travel to a location with community transmission of COVID-19, or, (ii) contact with a confirmed or probable COVID-19 case in the 14 days preceding symptom onset, or (iii) absence of an alternative diagnosis that completely explains the clinical presentation[6] was considered as suspected COVID-19 case.

Initially, a traveller was considered a suspected case only when arriving from China, although the definition of suspected cases associated with travel later included Japan, Singapore, South Korea, North Korea, Thailand, Vietnam and Cambodia (February 21, 2020), Italy, Germany, Australia, United Arab Emirates, Philippines, France, Iran and Malaysia (February 25, 2020), the USA, Canada, Switzerland, United Kingdom and 4 additional countries (March 3, 2020). From March 9, 2020 onwards, the Ministry of Health decided to start testing all hospitalised patients with severe respiratory symptoms, regardless of travel history.

Contact with a confirmed or probable COVID-19 case was defined as face-to-face or direct contact with a COVID-19 case, or direct contact in a health-care setting. Moreover, patients reporting travel to an affected country in the preceding 14 days were considered imported cases. Cases not meeting this criterion were considered to be due to local transmission.

Suspected COVID-19 cases were confirmed by laboratory testing (i.e., molecular diagnostic with real-time quantitative PCR), or by clinical-epidemiological criteria. In the latter case, the classification is used when laboratory testing is inconclusive or unavailable, as recommended by Brazilian Ministry of Health guidelines, dated April 6, 2020[49], and by the World Health Organization interim guidance, dated March 25, 2020[50].

Individual-level notification of COVID-19 and SARI cases with unknown aetiology from Brazil

To investigate individual-level diagnostic, demographic, self-reported travel history, place of residence and likely place of infection, differential diagnosis for other respiratory pathogens, as well as clinical details, including comorbidities, we collected three epidemiological data sources: (*i*) *n*= 67,344 suspected and *n*=1,468 confirmed cases notified to the REDCap database from February 25 to March 25, 2020; (*ii*) *n*=73,637 confirmed SIVEP-Gripe (*Sistema de Informação de Vigilância Epidemiológica da Gripe*) from March 1 to May 31, 2020 (available at https://shiny.hmg.saude.gov.br/dataset); and (*iii*) *n*=514,200 confirmed cases from aggregated data daily released at the *Portal do COVID-19* (Brazilian Health Ministry) from February 25 to May 31, 2020 (available at www.covid.saude.gov.br/). SIVEP-Gripe system notifies severe acute respiratory infections (SARI), which can be defined as an acute respiratory infection with onset within the last 10 days of fever (≥ 38º C) and cough, and typically requires hospitalization (see also **Fig. 1A**).

Basic reproduction number (R_0) estimation

We estimated the basic reproduction number (R_0) for SARS-CoV-2 using time series of confirmed COVID-19 cases at the national and state level: São Paulo, Rio de Janeiro,

Ceará and Amazonas (**Extended Data Fig. 1**). To avoid the impact of non-pharmaceutical interventions (NPI) on R_0 estimates, only data points up to 14 days after the implementation of the strictest interventions were used. As lockdown was not imposed in Brazil, the strictest measure was considered closure of non-essential commerce. For European countries, the date of lockdown was used as NPI date. NPI dates for Brazilian states were collected from state decrees. For Brazil as a whole the NPI date for São Paulo state was used, as by that point most states in Brazil had already closed non-essential commerce. For the European countries, lockdown dates were collected from https://www.covid19healthsystem.org/mainpage.aspx.

To test the estimation routine and provide international context, this analysis was replicated on equivalent time series from Italy, Spain, France, and the United Kingdom. Aggregated USA and China epidemiological data were not included due to possible heterogeneity within each country. Daily counts of confirmed cases were modelled with a negative binomial distribution with a mean equal to a fixed portion, ρ, of the total daily number of cases in an exponential model of incidence. The functional form of the incidence model is $\rho R_0 \gamma i_0 e^{(R0 - 1)\gamma t}$, which comes from an exponential approximation of the early dynamics where individuals cease to be infectious at a rate γ. The factor of $\rho R_0 \gamma$ accounts for the partial observation of the incidence. In this analysis was not accounted for the delay between infection and reporting.

Since ρ and i_0 only appear together, they were unidentifiable, we combine them into a single parameter, ξ. This identifiability issue prevents us from estimating the prevalence without additional information to inform either i_0 or ρ. The analysis was carried out in a Bayesian framework with an uninformative prior distribution on R_0 and an informative prior on the removal rate, all other parameters had weakly-informative prior distributions (details in the **Supplementary Information**, pp. 2-3). The informative prior ensured an individual is infectious for an average of 5 to 14 days[51] (**Supplementary Information**, **Fig. 5-6**). Standard

diagnostics were used to check whether the Markov Chain Monte Carlo (MCMC) samples were satisfactory. Full details of the model used, the estimation process and convergence of MCMC chains can be found in the **Supplementary Information**, pp. 2-3.

Geospatial analysis of COVID-19 cases and socio-economic status

The average household *per capita* income for the Metropolitan Region of São Paulo (MRSP) was retrieved at the census tract level from the 2010 census (https://censo2010.ibge.gov.br/). We geocoded 24,063 COVID-19 cases and 32,914 SARI cases with unknown aetiology from MRSP, which were notified until May 28, 2020. The geo-coding was based on self-reported residential address or postal codes using the Galileo algorithm[52] and coordinates were confirmed using the Google API.

To elucidate the distribution of COVID-19 cases and SARI cases with unknown aetiology cases, we mapped the mean relative risk of COVID-19 and SARI cases with unknown aetiology at the census tract level for MRSP for three epidemiological weeks (12, 16, and 21). As the observation process was a confounding process and without additional assumptions (e.g. covariates), we cannot disentangle an increase in prevalence from an increase in case ascertainment. The cumulative number of cases in each tract is modelled as a Poisson random variable with a mean specified by the expected number of cases under a null model adjusted by tract specific risk due to spatial and non-spatial effects: the Besag-York-Mollié model[19]. Estimates of the risk of COVID-19 diagnosis or SARI cases with unknown aetiology were obtained using approximate Bayesian methods (Integrated Nested Laplace Approximation). A complete specification of the model and the computational methodology can be found in the **Supplementary Information,** pp.1-2.

The association between final diagnostic category (COVID-19 or SARI cases with unknown aetiology) and socio-economic status in the subset of cases in the MRSP with geocoded residential information was evaluated using logistic regression models. We focused on the cases in epidemiological weeks 12, 16 and 22. Within each of those weeks, if a census tract reported any COVID-19 or SARI cases with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 of each census tract were mostly either 0 or 1 in a given week. For this reason, we defined two categories: (i) the census tract only reported SARI of unknown etiology, i.e. no COVID-19 cases, (ii) the census tract reported at least one COVID-19 case in the week. We used these two categories as the binary response, and applied logistic regression models to investigate whether income per capita was associated with this response. The analyses were adjusted by the logarithm of the population sizes and the longitude and latitude coordinates of the census tracts. The analysis was performed individually for each of epidemiological weeks 12, 16 and 22. Further details of this analysis can be found in the **Supplementary Information**, pp. 1-2.

Data availability

Datasets of clinical and laboratory data presented in the current study from SIVEP-Gripe and *Portal do COVID-19* database are available at https://doi.org/10.5061/dryad.n8pk0p2sp. The REDCap database and geolocation information are available from the corresponding authors upon request and ethical approval.

Code availability

The custom code used in this study is avaiable at https://doi.org/10.5061/dryad.n8pk0p2sp.

Author contributions

W.M.S, L.F.B, D.S.C, R.H.M.P, C.A.P, J.C, J-P.C, V.H.N, A.E.Z, J.M, F.C.S.S, P.S.A, F.G, A.A.S-S, B.G, C-H.W, S.L, N.G, S.B.O, K.V.P, M.C.T.D.B, V.B.G.P, C.K.V.B, F.G, W.A.F.A, F.F.S.T.F, E.M.M and W.K.O collected the epidemiological, spatial and clinical data and processed statistical data. N.R.F, W.M.S, L.F.B, C-H.W, J-P.C, D.C.S, R.H.M.P, J.M, E.C.S, P.M, S.L, L.A, A.A.S-S, G.L, A.T, M.F.V-G, M.U.G.K, R.S.A, N.A, P.M, O.J.B, I.O.M.S, N.G, G.L, O.G.P, A.E.Z, M.L.N, and J.C interpreted the results and wrote the manuscript. All authors read and revised the final manuscript. W.M.S, L.F.B, J.C, and N.R.F are responsible for summarising epidemiological and clinical data.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

The authors thank the clinicians and epidemiologists for technical support and Lucy Matkin for technical assistance. This work was supported by a FAPESP (2018/14389-0) and Medical Research Council and CADDE partnership award (MR/S0195/1), (http://caddecentre.org/). WMS is supported by the São Paulo Research Foundation, Brazil (No. 2017/13981-0 and 2019/24251-9). NRF is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). OJB was funded by a Sir Henry Wellcome Fellowship funded by the Wellcome Trust (206471/Z/17/Z). VHN and CAP were supported by FAPESP (2018/12579-7). AEZ and BG are supported by Oxford Martin School. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1 Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497-506, doi:10.1016/S0140-6736(20)30183-5 (2020).

2 Coronaviridae Study Group of the International Committee on Taxonomy of, V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, doi:10.1038/s41564-020-0695-z (2020).

3 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574, doi:10.1016/S0140-6736(20)30251-8 (2020).

4 Guan, W. J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *The New England Journal of Medicine*, doi:10.1056/NEJMoa2002032 (2020).

5 Livingston, E. & Bucher, K. Coronavirus Disease 2019 (COVID-19) in Italy. *Jama*, doi:10.1001/jama.2020.4344 (2020).

6 World-Health-Organization. Coronavirus disease (COVID-2019) situation reports (2020). Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

7 Croda, J. *et al.* COVID-19 in Brazil: advantages of a socialized unified health system and preparation to contain cases. *Revista da Sociedade Brasileira de Medicina Tropical* **53**, e20200167, doi:10.1590/0037-8682-0167-2020 (2020).

8 Jesus, J. G. *et al.* Importation and early local transmission of COVID-19 in Brazil, 2020. *Revista do Instituto de Medicina Tropical de Sao Paulo* **62**, e30, doi:10.1590/s1678-9946202062030 (2020).

9 Candido, D. S., Watts, A., Abade, L., Kraemer, M. U.G., Pybus, O. G., Croda, J., de Oliveira, W., Khan, K., Sabino, E. C., Faria, N. R. Routes for COVID-19 importation in Brazil. *Journal of Travel Medicine* **in press** (2020).

10 Clark, A., Jit, M., Warren-Gash, C., Guthrie, B., Wang, H. H. X., Mercer, S. W., Sanderson, C., McKee, M., Troeger, C., Ong, K. L., Checchi, F., Perel, P., Joseph, S., Gibbs, H. P., Banerjee, A., Eggo, R. M., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *The Lancet Global health* **https://doi.org/10.1016/ S2214-109X(20)30264-3** (2020).

11 Burki, T. COVID-19 in Latin America. *The Lancet Infectious Diseases* **20**, 547-548, doi:10.1016/S1473-3099(20)30303-0 (2020).

12 Cimerman, S., Chebabo, A., Cunha, C. A. D. & Rodriguez-Morales, A. J. Deep impact of COVID-19 in the healthcare of Latin America: the case of Brazil. *The Brazilian Journal of Infectious Diseases* **24**, 93-95, doi:10.1016/j.bjid.2020.04.005 (2020).

13 Ezequiel, G. *et al.* The COVID-19 Pandemic: A Call to Action for Health Systems in Latin America to Strengthen Quality of Care. *Int J Qual Health Care*, doi:10.1093/intqhc/mzaa062 (2020).

14 Miller, M. J., Loaiza, J. R., Takyar, A. & Gilman, R. H. COVID-19 in Latin America: Novel transmission dynamics for a global pandemic? *PLoS Negl Trop Dis* **14**, e0008265, doi:10.1371/journal.pntd.0008265 (2020).

15 Andrus, J. K. *et al.* Perspectives on Battling COVID-19 in Countries of Latin America and the Caribbean. *The American Journal of Tropical Medicine and Hygiene*, doi:10.4269/ajtmh.20-0571 (2020).

16 Croda, J. H. R. & Garcia, L. P. Immediate Health Surveillance Response to COVID-19 Epidemic. *Epidemiol Serv Saude* **29**, e2020002, doi:10.5123/S1679-49742020000100021 (2020).

17 Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform* **95**, 103208, doi:10.1016/j.jbi.2019.103208 (2019).

18 WHO. Influenza update (2020). https://www.who.int/influenza/surveillance_monitoring/updates/latest_update_GIP_surveillance/en/ (2020).

19 Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-20, doi:10.1007/BF00116466 (1991).

20 IBGE. Pesquisa nacional de saúde (2013) Brasil e grandes regiões. (2015).

21 Mellan, T. A. *et al.* Report 21: Estimating COVID-19 cases and reproduction number in Brazil. *medRxiv*, 2020.2005.2009.20096701, doi:10.1101/2020.05.09.20096701 (2020).

22 Caicedo-Ochoa, Y., Rebellon-Sanchez, D. E., Penaloza-Rallon, M., Cortes-Motta, H. F. & Mendez-Fandino, Y. R. Effective Reproductive Number estimation for initial stage of COVID-19 pandemic in Latin American Countries. *International Journal of Infectious Diseases* **95**, 316-318, doi:10.1016/j.ijid.2020.04.069 (2020).

23 Munayco, C. V. *et al.* Early transmission dynamics of COVID-19 in a southern hemisphere setting: Lima-Peru: February 29th-March 30th, 2020. *Infect Dis Model*, doi:10.1016/j.idm.2020.05.001 (2020).

24 Candido, D. d. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *medRxiv*, 2020.2006.2011.20128249, doi:10.1101/2020.06.11.20128249 (2020).

25 Ferguson, N. *et al.* Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Imperial College COVID-19 Response Team (2020).

26 Walker, P. G. T. *et al.* The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science*, doi:10.1126/science.abc0035 (2020).

27 Korber, B., Fischer, W. M., Gnanakaran,S., Yoon, H., Theilr, J., Abfalterer, W., Foley, B., Giorgi, E. E., Bhattacharya, T., Parker, M. D., Partrridgee, D. G., vans, C. M., Freeman, T. M., de Silva, T. I., on behald of the Sheffield. COVID-19 Genomics Group, LaBranche, C. C., Montefiori, D. C. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* **https://doi.org/10.1101/2020.04.29.069054** (2020).

28 Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*, 2020.2006.2012.148726, doi:10.1101/2020.06.12.148726 (2020).

29 Khalatbari-Soltani, S., Cumming, R. G., Delpierre, C. & Kelly-Irving, M. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J Epidemiol Community Health*, doi:10.1136/jech-2020-214297 (2020).

30 Rivett, L. *et al.* Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *eLife* **9**, doi:10.7554/eLife.58728 (2020).

31 Park, S. Y. *et al.* Coronavirus Disease Outbreak in Call Center, South Korea. *Emerging Infectious Diseases* **26**, doi:10.3201/eid2608.201274 (2020).

32 Pereira, R. H. et al. Mobilidade urbana e o acesso ao Sistema Único de Saúde para casos suspeitos e graves de Covid-19 nas vinte maiores cidades do Brasil. 14 Diretoria de Estudos e Politicas Regionais, Urbanas e Ambientais, IPEA (2020).

33 Silveira, M. *et al.* Repeated population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. *medRxiv*, 2020.2005.2001.20087205, doi:10.1101/2020.05.01.20087205 (2020).

34 Sy, K. T. L., Martinez, M. E., Rader, B. & White, L. F. Socioeconomic disparities in subway use and COVID-19 outcomes in New York City. *medRxiv*, 2020.2005.2028.20115949, doi:10.1101/2020.05.28.20115949 (2020).

35 Dehning, J. *et al.* Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, doi:10.1126/science.abb9789 (2020).

36 Buckee, C. O. *et al.* Aggregated mobility data could help fight COVID-19. *Science* **368**, 145-146, doi:10.1126/science.abb8021 (2020).

37 de Oliveira, S. B. *et al.* Monitoring social distancing and SARS-CoV-2 transmission in Brazil using cell phone mobility data. *medRxiv*, 2020.2004.2030.20082172, doi:10.1101/2020.04.30.20082172 (2020).

38 Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493-497, doi:10.1126/science.abb4218 (2020).

39 Nouvellet, P. *et al.* Report 26: Reduction in mobility and COVID-19 transmission. Imperial College COVID-19 Response Team (2020).

40 Wu, X. *et al.* Co-infection with SARS-CoV-2 and Influenza A Virus in Patient with Pneumonia, China. *Emerging Infectious Diseases* **26**, doi:10.3201/eid2606.200299 (2020).

41 Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. *JAMA*, doi:10.1001/jama.2020.6266 (2020).

42 Cuadrado-Payan, E. *et al.* SARS-CoV-2 and influenza virus co-infection. *Lancet* **395**, e84, doi:10.1016/S0140-6736(20)31052-7 (2020).

43 Wu, X. *et al.* Co-infection with SARS-CoV-2 and Influenza A Virus in Patient with Pneumonia, China. *Emerging Infectious Diseases* **26**, 1324-1326, doi:10.3201/eid2606.200299 (2020).

44 Zheng, X. *et al.* Co-infection of SARS-CoV-2 and Influenza virus in Early Stage of the COVID-19 Epidemic in Wuhan, China. *The Journal of infection*, doi:10.1016/j.jinf.2020.05.041 (2020).

45 Asner, S. A. *et al.* Clinical disease severity of respiratory viral co-infection versus single viral infection: a systematic review and meta-analysis. *PloS One* **9**, e99392, doi:10.1371/journal.pone.0099392 (2014).

46 Black, A., MacCannell, D. R., Sibley, T. R. & Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine*, doi:10.1038/s41591-020-0935-z (2020).

47 Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science*, doi:10.1126/science.abb9263 (2020).

48 Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003 e1009, doi:10.1016/j.cell.2020.04.023 (2020).

49 Ministerio da Saude, B. Coronavirus COVID-19 Diretrizes para Diagnostico e Tratamento da COVID-19. (2020).

50 WHO. *COVID-19 coding in ICD-10*, <https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf?ua=1> (2020).

51 Wolfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature*, doi:10.1038/s41586-020-2196-x (2020).

52 Medel, C. H., Catalan, C. C. Vidou, M. A. F., Perez, E. S. The Galileo Ground Segment Integrity Algorithms: Design and Performance. *International Journal of Navigation and Observation* **2008** (2008).

Legend figures

Fig. 1 | Timeline of national COVID-19 notification systems in Brazil. The REDCap system operated between late January until March 25, 2020. Aggregated numbers from e-SUS-VE and SIVEP-Gripe data for mild and hospitalised COVID-19 cases, respectively, are updated on a daily basis at the *Portal do COVID-19* website (https://covid.saude.gov.br/).

Fig. 2 | COVID-19 epidemiology in Brazil. a. Number of COVID-19 cases (blue filled line) and deaths (blue dashed line) reported to the Ministry of Health (*Portal do COVID-19* website), and number of COVID-19 confirmed cases (salmon filled line) and number of SARI with unknown aetiology (salmon dashed line) reported to the SIVEP-Gripe database. **b**. First COVID-19 cases by date and Brazilian municipal population size based on the Ministry of Health, from March 28, 2020. Each circle represents the first confirmed COVID-19 case in the municipality (n= 4,196 Brazilian municipalities). **c**. Map coloured according to the number of confirmed COVID-19 cases per state reported to the Ministry of Health (*Portal do COVID-19* website). Circle sizes are proportional to the number of reported COVID-19 deaths in each federal unit. SPBR1 is the first detected SARS-CoV-2 infection in Brazil [8]. The codes for the 27 federal units in Brazil were: Acre (AC), Alagoas (AL), Amapá (AP), Amazonas (AM), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Mato Grosso (MT), Mato Grosso do Sul (MS), Minas Gerais (MG), Pará (PA), Paraíba (PB), Paraná (PR), Pernambuco (PE), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Rondônia (RO), Roraima (RR), Santa Catarina (SC), São Paulo (SP), Sergipe (SE) and Tocantins (TC).

Fig.3 | Estimated R_0 values for four Brazilian states and selected countries. Left, R_0 for the Amazonas, Ceará, Rio de Janeiro and São Paulo states. Right, R_0 for Brazil, France, Italy, Spain and United Kingdom. Daily number of infections used in each analysis can be found in **Extended Figs. 3-4**. Daily number of infections and prior distributions can be found in **Extended Figs. 5-6**.

Fig. 4 | COVID-19, SARI with unknown aetiology and influenza. Red and orange lines indicate cases notified in 2020, blue lines indicate cases notified in 2016 for influenza (filled blue line) and SARI cases with unknown aetiology (dashed blue line). Grey lines indicate influenza and SARI cases with unknown aetiology for 2017, 2018 and 2019.

Fig. 5 | COVID-19 diagnosis and socio-economic factors in the Metropolitan Region of São Paulo. A. Spatial distribution of income per capita of MRSP based on census tract of residence. **B.** Distribution of household *per capita* income based on census tract of residence for COVID-19 cases and SARI cases with unknown aetiology. The distribution of average *per capita* income for MRSP as a whole, weighted by population size, is shown on the left. **C.** Posterior mean relative risk of COVID-19 confirmed diagnosis (upper panels) and SARI cases with unknown aetiology (lower panels) for epidemiological weeks 12 (pre-implementation of NPI in São Paulo state, and weeks 16 and 21 (post-implementation of NPI in São Paulo state) (see **Methods** for details).

Fig. 6 | Age-sex structure and clinical features of confirmed COVID-19 cases notified on the SIVEP-Gripe system. A. Age classes are shown on the left of the panel. On-going cases were those still active on the SIVEP-Gripe database and without a recorded clinical outcome

761 (death or recovered). **B.** Symptoms, signs and comorbidities of confirmed COVID-19 cases.
762 **C**. Comorbidities among confirmed COVID-19 cases according to age groups and outcome.
763 Confirmed COVID-19 cases with complete comorbidity and outcome (death or recovery)
764 information (n = 15,720). Confirmed COVID-19 cases with complete information on
765 comorbidities and ITU admission (n = 19,409). Horizontal axes show the proportion of
766 patients in each age/outcome stratified with each of the comorbidities recorded.
767

Supplementary information

Geospatial analysis

We adopted a Bayesian hierarchical model to compute relative risk for each census tract, due to the following reasons: (i) there is a large number of census tracts (*n*=30,815), (ii) there is substantial heterogeneity in the size of census tracts, and (iii) small counts in each tract obscure the spatial distribution of observed cases. The number of observed cases in census tract i is modelled using a Poisson distribution $Y_i = Poisson(\lambda_i)$ with mean $\lambda_i = E_i\,\mu_i$ where E_i is the expected number of cases under a null model in which cases are uniformly distributed among the population. For example, the total number of cases in the MRSP multiplied by the proportion of the population in the census tract $E_{it} = \frac{\sum_i Y_i}{\sum_i pop_i} \times pop_i$. The factor of-$\mu_i$ describes tract specific risk and models the additional variation in the observation process[1]. A log-linear model is used to estimate the relative risk μ_i. For example, the log relative risk is expressed as a sum of an intercept α, which represents the overall relative risk (in our case, the global relative risk is zero), and random effects (Z_i):

$$\log(\mu_i) = \alpha + Z_i$$

We used a Besag-York-Mollié model (BYM)[2] to separate the random effects into a spatially structured U_i, and independent random effects, V_i, so ($Z_i = U_i + V_i$). In the BYM model, a conditional autoregressive (CAR) process is used to introduce correlation among the U_i for each tract. Given the U_i of neighbouring tracts, the U_i has a normal distribution with mean equal to the average of the neighbours' U_i, and variance $s_i^2 = \frac{1}{\#N(i)\tau_U}$ where $\#N(i)$ is the number of tracts that share boundaries with tract i and τ_U is a precision parameter. The random effect, V_i follows a zero mean normal distribution with unknown precision, $\tau_V = \frac{1}{\sigma_{v^2}}$ (where σ_{V^2} is the variance). Both random effects in the model capture extra-Poisson variability, and were expressed as the following:

$$U_i | U_{j\neq i} \sim Normal(m_i, s_i^2), \quad V_i \sim N(0, \sigma_v^2)$$

$$m_i = \frac{\sum_{j\in N(i)} U_j}{\#N(i)}, \quad s_i^2 = \frac{\sigma_U^2}{\#N(i)} = \frac{1}{\#N(i)\tau_U}$$

The log of the precision parameters, τ_U and τ_V, follows a gamma distribution with shape 1 and rate 0.0005. These are the default priors used by R-INLA and are minimally informative[3]. The prior default distributions in R-INLA were used for the precision parameters of both U_i and V_i. These are minimally informative and are the recommended settings [4].

To quantify the uncertainty in the point estimates of the mean relative risk estimates, we mapped the posterior probability of elevated relative risk in each census tract (**Extended Data Fig. 9**). This is the posterior probability, which a tract has an elevated risk of observing cases, formally, this is Prob(μ_i > 1| data). For instance, a probability of 0.6 in a census tract indicates a 60% chance that this census tract is at greater risk of observing cases relative to the rest of the MRSP.

Analysis of the relationship between income per capita and final diagnostic category in the Metropolitan Region of Sao Paulo (MRSP)

We evaluated the relationship between final diagnostic category (COVID-19 or SARI cases with unknown aetiology) and socioeconomic status in the subset of cases in the MRSP with geocoded residential information. We focused on the cases in epidemiological weeks 12, 16 and 22, where the census tracts that reported cases varied across weeks. In each of the three weeks, if a census tract reported any COVID-19 or SARI cases with unknown aetiology with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 of each census tract were mostly either 0 or 1 in a given week. Based on this observation and let i index the census tracts, we subsequently defined the binary outcome Y_i of census tract i, where (i) $Y_i = 0$ if census tract i only reported SARI cases with unknown aetiology with unknown aetiology, i.e. no COVID-19 cases, (ii) $Y_i = 1$ if census tract i reported at least one COVID-19 case in the week. Logistic regression models were applied to investigate the association between this binary outcome and the log(X+1) transformed income per capita. The analyses were adjusted by the logarithm of the population sizes. In addition, the census tracts were grouped by their geographic locations using cluster analysis, and the groupings were used as the random effect in the logistic regressions to account for potential spatial autocorrelation. The number of clusters was chosen based on the AIC/BIC values of the logistic regression models. The analysis was performed individually for each of epidemiological weeks 12, 16 and 22.

A likelihood ratio test (LRT) is applied to each analysis to examine whether the log(X+1) transformed income per capita provides information in addition to the information from the log population size and the random effects. The regression coefficients and LRT P-values of income are presented in (**Supplementary Table S3**).

Estimating basic reproduction number (R_0)

Since SARS-CoV-2 is a novel virus, and we are subsetting data to avoid the impact of either non-pharmaceutical interventions or depletion of the susceptible pool, we deemed it reasonable to model the incidence of infection with an exponential approximation to the early behaviour of an SIR model, i.e., the incidence grows exponentially [5]. This model makes several strong assumptions about the dynamics of the epidemic: (i) the populations under consideration mix homogeneously, (ii) the proportion of the population that is susceptible stays close to 100%, (ii) the proportion of infections that are observed, and the basic reproduction number are constant throughout time, and (iv) the delay between infection, and notification is a constant. Although there are obvious violations of these assumptions, they provide a convenient starting point for estimating the basic reproduction number. Ignoring the delay between infection and observation will on average only translate the results in time by the incubation period and the delay from infection to diagnosis.

Under the assumptions outlined above, the expected number of daily cases, $\mu(n)$ on day n is given by the following equation: $\mu(n) = \rho \mathrm{R}_0 \gamma i_0 e^{(\mathrm{R}0-1)\gamma.n}$ where ρ is the probability of an infection being counted in the time series, R_0, is the basic reproduction number, γ is the rate at which individuals cease to be infectious and i_0, is the proportion of the population that was infectious at the start of the observations. We assume that the observed number of cases on day n was drawn from a negative binomial observation where the mean is $\mu(n)$ and the variance, $\sigma = \mu + \mu^2/k$, with fixed size parameter, k (*dispersion parameter*). The product of ρ and i_0 is denoted ξ. Since the probability of being observed and the initial condition only appear as the product ξ in the likelihood, there is an identifiability problem preventing the estimation of ρ and i_0 individually, consequently we only

consider their product, ξ. Although in this model it is theoretically possible to estimate both R_0 and γ, in practice this is difficult so we will use an informative prior to constrain γ to a priori plausible values.

Regarding prior distributions, for R_0 we used a uniform prior over values from 1 to 10. The removal rate, γ, was given an informative prior distribution: a normal distribution with mean (1/5 + 1/14) / 2 = 0.1357, leading to an average duration 7.4 days during which an individual is infectious. Moreover, the average duration of infectivity is constrained to be between the extremes of 5 and 14 days. These values for the infective duration were found in the literature [6,7]. The standard deviation of the prior distribution for γ is (1/5 - 1/14) / 4 = 0.03124, this ensures that 95% of the prior probability lay within these bounds. For the parameter ξ, we used a log-normal prior with a log mean of 0.0 and a log standard deviation of 1.0. For the size parameter of the negative binomial, k, a log-normal distribution was used with a log-mean of 0.0 and log-standard deviation of 1.0 to enable this parameter to have a large range of values.

Samples from the posterior distribution were obtained using MCMC running 4 chains from random initial conditions using the mcmc library available on CRAN2 and using coda for diagnostics[8,9]. Trace plots of the posterior samples suggested that the chain had converged and mixed, and there was an effective size of at least several hundred for each of the 4 parameters of this model. The prior and posterior distributions were checked to ensure that (beyond the removal rate) each parameter was being informed by the data. Each data set: Brazil and Europena countries (Italy, the United Kingdom, France, and Spain) or Brazilian states (São Paulo, Rio de Janeiro, Amazonas, and Ceará) were run as independent analyses, the model fit from the point estimate along with the corresponding trace plots and prior/posterior comparisons is shown in **Extended Data Figs. 5 and 6.**

References

1 Lawson, A. B. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiolog*. (2008).

2 Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-20, doi:10.1007/BF00116466 (1991).

3 Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319-392, doi:10.1111/j.1467-9868.2008.00700.x (2009).

4 Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology* **7**, 39-55, doi:10.1016/j.sste.2013.07.003 (2013).

5 Brauer, F., van den Driessche, P. & Wu, J. *Mathematical Epidemiology*. (Springer-Verlag Berlin Heidelberg, 2008).

6 Wolfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature*, doi:10.1038/s41586-020-2196-x (2020).

7 European Centre for Disease Prevention and Control. Novel coronavirus (SARS-CoV-2). (2020).

8 Geyer, C.J., & Jonson, L. T. mcmc: Markov Chain Monte Carlo. R package version 0.9-6 (https://CRAN.R-project.org/package=mcmc, 2019).

9 Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7-11 (2006).