Title: Model-based estimation of transmissibility and reinfection of SARS-CoV-2 P.1 variant

Authors: Renato Mendes Coutinho[1,7], Flavia Maria Darcie Marquitti[2,7], Leonardo Souto Ferreira[3,7], Marcelo Eduardo Borges[7], Rafael Lopes Paixão da Silva[3,7], Otavio Canton[3,6], Tatiana P. Portella[4,7], Silas Poloni Lyra[3,7], Caroline Franco[3,7], Antonio Augusto Moura da Silva[6,7], Roberto A. Kraenkel[3,7], Maria Amélia de Sousa Mascena Veras[5,7], Paulo Inácio Prado[4,7]

Affiliations:
1. Universidade Federal do ABC, Santo André, SP, Brazil
2. Universidade Estadual de Campinas, Campinas, SP, Brazil
3. Universidade Estadual Paulista, São Paulo, SP, Brazil
4. Universidade de São Paulo, São Paulo, SP, Brazil
5. Faculdade de Ciências Médicas da Santa Casa de São Paulo – São Paulo, SP, Brazil
6. Universidade Federal do Maranhão, São Luís, MA, Brazil
7. Observatório COVID-19 BR

Abstract

The variant of concern (VOC) P.1 emerged in the Amazonas state (Brazil) and was sequenced for the first time on 6-Jan-2021 by the Japanese National Institute of Infectious Diseases. It contains a constellation of mutations, ten of them in the spike protein. Consequences of these mutations at the populational level have been little studied so far. From December-2020 to February-2021, Manaus was devastated by four times more cases compared to the previous peak (April-2020). Here, data from the national health surveillance of hospitalized individuals and frequency of the P.1 variant were analysed using a model-based approach to estimate P.1 parameters of transmissibility and reinfection by maximum likelihood. Sensitivity analysis was performed changing pathogenicity and the period analysed (including/excluding the health system collapse period). The new variant transmissibility was found to be about 2.5 times higher (Confidence Interval (95%CI): 2.3–2.8) compared to the previous variant in Manaus. A low probability of reinfection by the new variant (6.4%, 95%CI: 5.7–7.1%).) was estimated, even under initial high prevalence (68%, 95%CI: 63–74%), by the time P.1 emerged. Consequences of a higher transmissibility were already observed with VOC B.1.1.7 in the UK and Europe. Urgent measures must be taken to control the spread of P.1.

Introduction

The Japanese National Institute of Infectious Diseases identified the new P.1 SARS-CoV-2 variant from travelers returning from Amazonas State, Brazil on 6-Jan-2021 (1). P.1 was first reported in Manaus, the Amazonas state capital city, on 11-Jan-2021 (2). Later, it was identified in samples collected since 6-Dec-2020 from Manaus (3). According to phylogenetic studies, P.1 likely emerged in the Amazonas state in early (3) or late (4) November-2020. This variant contains a constellation of mutations, ten of them in the spike protein, and shares mutations with other variants of concern (VOCs) previously detected in the United Kingdom and South Africa (B.1.1.7 and B.1.351, respectively), including E484K, K417T, N501Y, and a deletion in the orf1b protein (del11288-11296 (3675-3677 SGF)) (2).

The Coronavirus disease 2019 (COVID-19) outbreak in Manaus (April-May 2020) was followed by a period of high but stable incidence. Then, from December 2020 to February 2021 the city was devastated by a new outbreak that caused a collapse in the health system and shortages of oxygen supply. (5). The frequency of P.1 increased sharply from 0% in November 2020 to 73% in January 2021 (4). The pathogenicity of the P.1 variant is still unknown, although recent studies point to increased viral load in individuals infected with the new variant (4). We analyzed national health surveillance data of hospitalizations and frequency of P.1 sequences from residents of Manaus city using a model-based approach to estimate P.1 transmissibility and reinfection.

The study

Dataset

We used the Brazilian epidemiological syndromic surveillance system for influenza, SIVEP-Gripe (publicly available at https://opendatasus.saude.gov.br), to track COVID-19 hospitalized cases. All hospitalized patients with Severe Acute Respiratory Infections are reported to SIVEP-Gripe with symptom onset date and SARS-CoV-2 test results. Data for hospitalized COVID-19 cases among residents in Manaus from 01-Nov-2020 to 31-Jan-2021 was obtained from SIVEP-Gripe database from 15-Feb-2021. The hospitalized cases of the last 10 weeks in the time series were nowcasted (6) to correct for notification delay. Time-series of frequency of sequenced genomes identified as P.1 in Manaus were extracted from published datasets (3,7).

Model

A deterministic compartmental model (Figure 1) was developed to model the infection of *Susceptible* individuals moving to the *Exposed* (pre-symptomatic) compartment, which can progress to three alternative compartments: *Hospitalized* (severely ill), *Infected*

(symptomatic but non-hospitalized), and *Asymptomatic*. Eventually, individuals move to *Recovered* or *Deceased*. Two variants are considered: 1-non-P.1 ("wild-type") and 2-new/P.1. The latter is assumed to infect *Recovered* individuals previously infected by the wild-type, and no reinfections of wild-type due to waning immunity occur. Compartments were stratified into three age categories: young (<20 years old), adult (≥20 and <60 years old) and elderly (≥60 years old), with different rates for outcomes. The key parameters of *relative transmissibility* and *reinfection probability* were estimated by a maximum likelihood fitting to the weekly number of new hospitalizations and to genomic surveillance data (3,7). Three additional model parameters with unknown values were estimated. The remaining parameters (24 out of 29) were fixed , using current values from the literature (see Table A.I in the Appendix for values and references). Sensitivity to higher pathogenicity of the variant P.1 was explored by repeating the fit assuming higher infection hospitalization rate (SA1). The sensitivity to the period analysed was also explored by another fit excluding the health system collapse period (SA2). Further details of the model and fitting methodology are available in the Appendix.

Results

The estimated transmissibility of P.1 was 2.5 (95% Confidence Interval (CI): 2.3–2.8) times higher compared to the wild-type variant, while the reinfection probability due to the new variant was 6.4% (CI:5.7–7.1%). The model fitting also estimated that, at the time the new P.1 variant emerged, the prevalence of previous infection to the wild-type variant was 68% (CI:63–74%), and that the number of cases by the wild-type variant were increasing with an estimated daily intrinsic growth rate of 0.053 $days^{-1}$ (CI:0.047–0.058 $days^{-1}$). See Table 1 and Figure 2 for fitted parameters results. The results were robust and confirmed by sensitivity analyses as shown in Table 2.

Conclusions

COVID-19 hospitalizations and frequency of the P.1 variant in clinical samples showed a sharp increase in Manaus, Brazil, starting November 2020. The fitted model suggests this joint increase was the result of the emergence of P.1, estimated to be 2.5 times more transmissible than the wild-type variant. The spread of P.1 occurred despite an existing high estimated prevalence of infection by the wild-type virus and a small probability of reinfection by the invading variant. The pathogenicity of P.1 is still unknown, but even assuming higher infection hospitalization rates for the P.1 variant, the estimated transmissibility remained above twice that of the wild-type variant.

Two recent studies analysed genomic data of SARS-CoV-2 from Manaus evaluating the transmissibility of the new variant (3,4). Faria and collaborators integrated mortality and genomic data and, using a semi-mechanistic Bayesian model, estimated a transmissibility

1.4–2.2 times higher and 25–61% evasion of protective immunity related to the variant P.1 (3). Naveca and collaborators estimated a 2.2 times higher effective reproduction number for the P.1 variant using phylogenetic methods, and suggested that P.1 is at least two times more transmissible than the parental lineage, assuming reinfections are rare (4). The present work follows a different approach that can be defined as an epidemiological, model-based, and data-fitting approach. Notably, the three different approaches all estimated higher transmissibility of the P.1 variant.

Data from the Amazonian region are scarce. The existing prevalence data were obtained from convenience samples, and P.1 frequency over time was based on small sample sizes; the former was included in the parameters fitted, while the uncertainty of the latter was explicitly considered. Epidemiological data are probably incomplete, since the analysed data overlap with the period of the health system collapse, posing limitations to this study. This was minimized by removing the collapse period in the sensitivity analysis (SA2) and the results were still robust. Due to lack of data evaluating waning immunity, the present work does not include reinfection by the wild-type on recovered individuals. If waning immunity is an important aspect, the transmissibility might be lower than estimated in this study. However, a recent study considering this possibility still estimated a high transmissibility for the P.1 variant (3).

The threats of a highly transmissible variant have already been observed with VOC B.1.1.7 in the UK, USA and Europe (8). Higher transmissibility of the P.1 variant raises strong concerns of swift upsurges in the number of cases once P.1 reaches community transmission across Brazil and other countries. The P.1 variant has already been detected in at least 25 countries (8). This calls for urgent i) pathogenicity studies of the P.1 variant, since greater transmissibility and pathogenicity can drive even well-prepared health systems to collapse; ii) immunological studies to understand the role of waning immunity in emergence of variants after a long period of epidemic control; and finally iii) mitigation measures to control the spread of P.1.

Acknowledgments: We are grateful for the collaborative work of the entire group of the Observatório COVID-19 BR, which has been tirelessly studying the situation of the SARS-CoV-2 epidemic in Brazil. The authors also thank the research funding agencies: the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (Finance Code 001 to FMDM, LSF and TPP), Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brazil (grant number: 315854/2020-0 to MEB, 141698/2018-7 to RLPS, 313055/2020-3 to PIP, 312559/2020-8 to MASMV, 311832/2017-2 to RAK, 305703/2019-6 to AAMS) and Fundação de Amparo à Pesquisa do Estado de São Paulo- Brazil (grant number: 2019/26310-2 and 2017/26770-8 to CF, 2018/26512-1 to OC, 2018/24037-4 to SPL and contract number: 2016/01343-7 to RAK).

References

1. Japan.; NIID [Internet]. Brief report: New Variant Strain of SARS-CoV-2 Identified in Travelers from Brazil. Coronavirus disease 4. [cited 2021 Feb 28]. Available from: https://www.niid.go.jp/niid/en/2019-ncov-e/10108-covid19-33-en.html
2. Naveca F, Nascimento V, Souza V, Corado A, Nascimento F, Silva G et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. Virological. 2021 [Preprint]. [cited 2021 Feb 28]. Available from: https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585 .
3. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, et al. Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. [Preprint]. 2021 [cited 2021 Mar 03]. Available from: https://doi.org/10.1101/2021.02.26.21252554
4. Naveca F, Nascimento V, Nascimento V, Souza V, Corado A, Nascimento F, Silva G, Costa A, et al. COVID-19 epidemic in the Brazilian state of Amazonas was driven by long-term persistence of endemic SARS-CoV-2 lineages and the recent emergence of the new Variant of Concern P.1. Research Square [Preprint]. 2021 [cited 2021 Feb 28]. Available from: https://doi.org/10.21203/rs.3.rs-275494/v1 .
5. Taylor L. Covid-19: Is Manaus the final nail in the coffin for natural herd immunity? BMJ. 2021;372:n394. Available from: http://www.ncbi.nlm.nih.gov/pubmed/33579721
6. McGough SF, Johansson MA, Lipsitch M, Menzies NA. Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. PLoS computational biology. 2020;16(4):e1007735.
7. Rede Genômica Fiocruz [Internet]. Plots of lineages presence by state. [cited 2021 Feb 28]. Available from: http://www.genomahcov.fiocruz.br/presenca-das-linhagens-por-estado/
8. PANGO lineages [Internet]. Global report investigating novel coronavirus haplotypes-GRINCH. [cited 2021 Mar 04]. Available from: https://cov-lineages.org/global_report_P.1.html

Appendix references

9. K. Soetaert, T. Petzoldt, and R. W. Setzer. Solving differential equations in R: Package deSolve. Journal of Statistical Software, 33(9):1–25, 2010. ISSN 1548- 7660. Available from: http://www.jstatsoft.org/v33/i09.

10. Y. Wei, L. Wei, Y. Liu, L. Huang, S. Shen, R. Zhang, J. Chen, Y. Zhao, H. Shen, and F. Chen. A systematic review and meta-analysis reveals long and dispersive incubation period of covid-19. medRxiv, 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.06.20.20134387v1
11. M. Cevik, M. Tate, O. Lloyd, A. E. Maraolo, J. Schafers, and A. Ho. SARS-CoV- 2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. The Lancet Microbe, 2(1):e13–e22, jan 2021. ISSN 26665247. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666524720301725
12. SIVEP-Gripe: Brazilian epidemiological syndromic surveillance system for influenza [Internet]. Database from OpenDataSUS. 2021. Available from: https://opendatasus.saude.gov.br/dataset?tags=SRAG
13. S. M. de Saúde Município de São Paulo. Inquérito sorológico para Sars-Cov-2: Prevalência da infecção em escolares das redes públicas e privada da cidade de São Paulo. 2021. [Online; accessed 31-January-2021]. Available from: http://www.capital.sp.gov.br/arquivos/pdf/2021/coletiva saude 14-01.pdf
14. W. W. Sun, F. Ling, J. R. Pan, J. Cai, Z. P. Miao, S. L. Liu, W. Cheng, and E. F. Chen. Epidemiological characteristics of COVID-19 family clustering in Zhejiang Province. Chinese journal of preventive medicine, 54(6):625–629, 2020. ISSN 02539624. Availbale from: https://doi.org/10.3760/cma.j.cn112150-20200227-00199
15. H. Salje, C. T. Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, et al. Estimating the burden of SARS-CoV-2 in France. Science, 369(6500): 208–211, May 2020. Available from: https://doi.org/10.1126/science.abc3517.
16. T. P. Portella, S. R. Mortara, R. Lopes, A. Sánchez-Tapia, M. R. Donalísio, M. C. Castro, et al. Temporal and geographical variation of COVID-19 in-hospital fatality rate in Brazil. meRxiv. 2021. Available from: https://www.medrxiv.org/content/10.1101/2021.02.19.21251949v1
17. K. Prem, K. van Zandvoort, P. Klepac, R. M. Eggo, N. G. Davies, A. R. Cook, M. Jit, et al. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the covid-19 era. medRxiv. 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.07.22.20159772v2
18. Y. Qiu, J. Mei, and M. Y. Qiu. Package rARPACK: Solvers for Large Scale Eigenvalue and SVD Problems. 2016. R package version 0.11-0. Available from: https://CRAN.R-project.org/package=rARPACK.
19. B. Bolker and R Development Core Team. bbmle: Tools for General Maximum Likelihood Estimation, 2020. R package version 1.0.23.1. Available from: https://CRAN.R-project.org/package=bbmle.

Figures and table

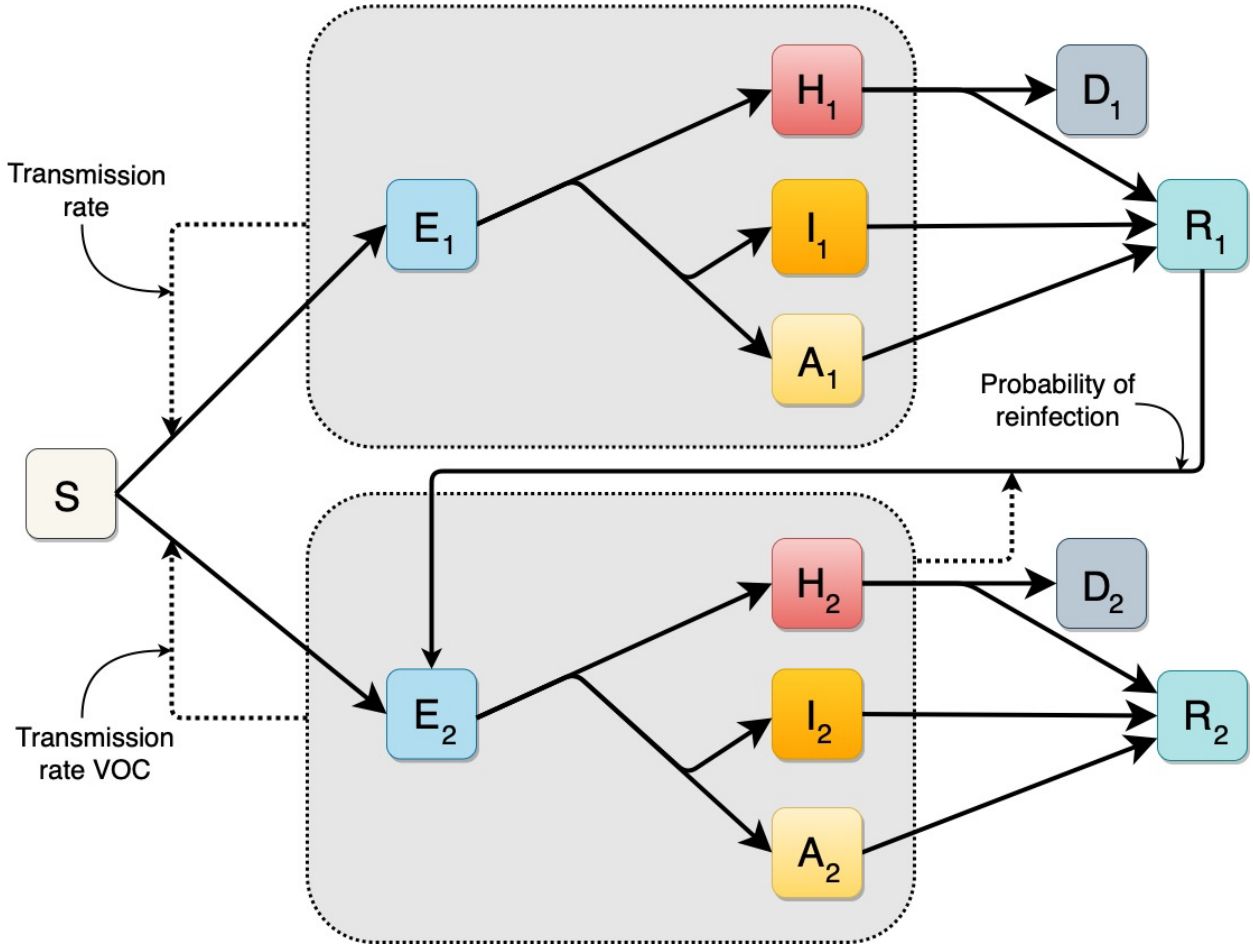

Figure 1: Diagram extended Susceptible, Exposed, Infected, and Recovered (SEIR) deterministic compartmental (SEAIHRD) model. S: Susceptible, E: Exposed (pre-symptomatic), H: Hospitalized (severe infected individuals), I: Infected (symptomatic individuals, not hospitalized), A: Asymptomatic. D: Deceased, R: Recovered. Compartments are subdivided into 3 age classes, not represented here for simplicity. Compartments with no subindex are related to the wild-type variant, compartments with subindex 2 are related to the VOC P.1. Continuous lines represent flux between each compartment. Dashed lines represent infection pathways. Three parameters fitted by the model are indicated by the small arrows.

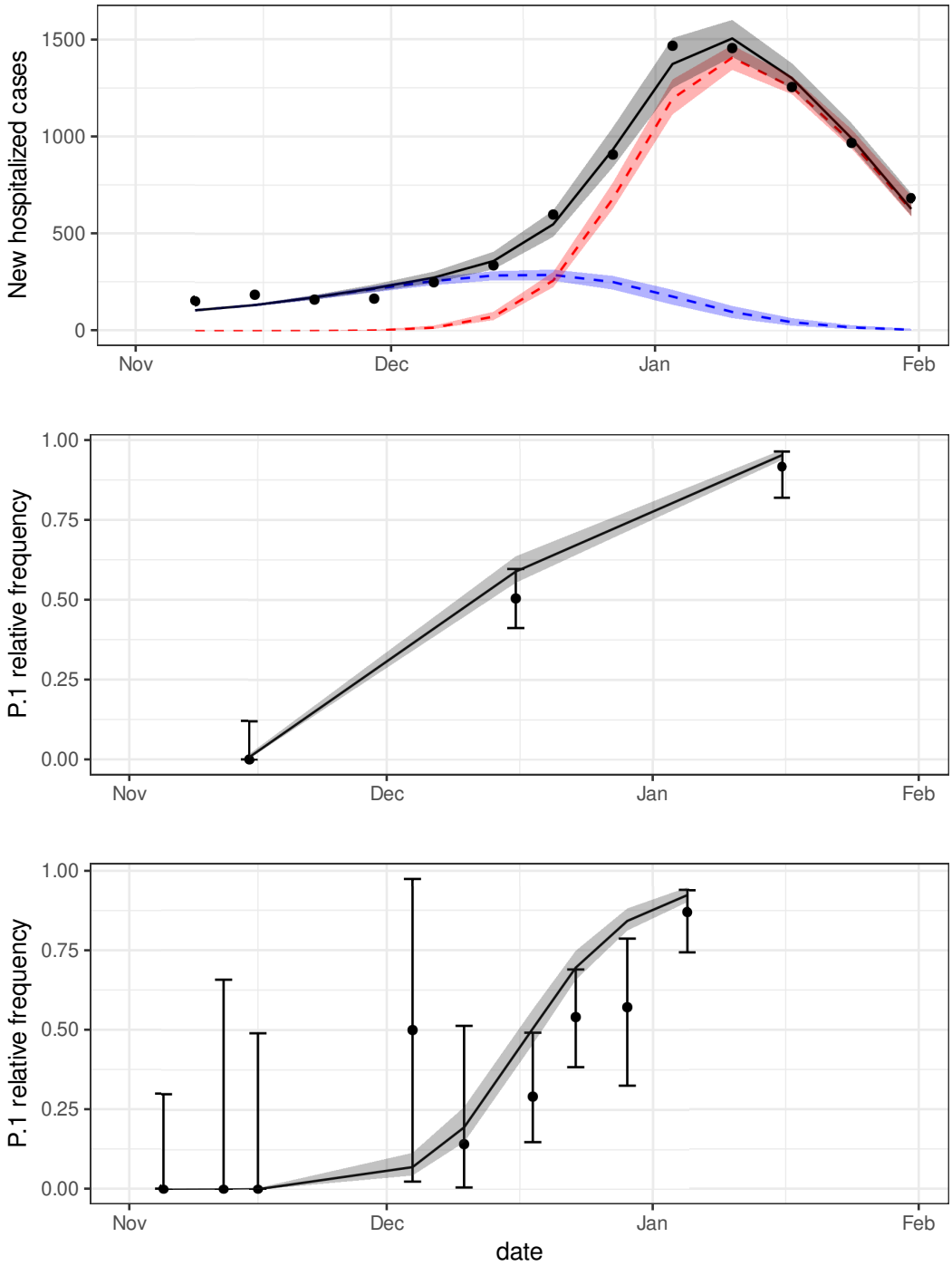

Figure 2: First panel: Weekly new hospitalized COVID-19 cases in Manaus city. Grey line represents the fitted values of total cases (all variants) by maximum likelihood estimation (MLE) of the parameters. Red and blue lines represent the fitted values of cases due to P.1 and wild-type variants, respectively. Black dots are nowcasted observed data of hospitalization. Second and third panels show the fittings to the P.1 time-series frequency of P.1 on datasets provided by (7) and (3) respectively. The area around the lines indicate the 95% Confidence Interval (CI) of the expected values. The fitted values of the model parameters are presented in the first part of Table 1.

Table 1: Summary of the fitted parameters and respective confidence intervals considering the entire period, November 1 2020-January 31, 2021 maintaining the same pathogenicity of the previous variant.

Parameter	estimate	2.50%	97.50%
Relative transmission rate for the new variant	2.52	2.27	2.78
Probability of reinfection by the new variant (%)	6.4	5.7	7.1
Prevalence of previous infection when the new variant emerges (%)	68	63	74
Initial fraction of the new variant (2020-11-01) (x 10^{-5})	3.7	0.6	23.4
Intrinsic growth rate ($days^{-1}$)	0.053	0.047	0.058

Table 2: Sensitivity analysis considering greater pathogenicity of the P.1 variant (SA1) and data censuring after the collapse of the healthcare system (SA2) in Manaus, Brazil, on 10-Jan-2021.

Parameter	estimate		2.50%		97.50%	
Sensitivity Analysis	**SA1**	**SA2**	**SA1**	**SA2**	**SA1**	**SA2**
Relative transmission rate for the new variant	2.56	2.81	2.33	2.50	2.78	3.11
Probability of reinfection by the new variant (%)	2.0	0.000	0.2	0.000	19.0	0.000
Prevalence of previous infection when the new variant emerges (%)	68	54	49	50	82	59
Initial fraction of the new variant (2020-11-01) (x 10^{-5})	10.2	5.9	0.6	1.1	174.2	33.1
Intrinsic growth rate (r)	0.050	0.046	0.038	0.043	0.062	0.049

Appendix

In order to estimate key parameters of the variant of concern (VOC) P.1, we developed a model and fitted it to time-series data on the number of hospitalized cases and frequency of the P.1 variant. Section I describes the model, section II relates the values of the parameters taken from the current literature, section III describes the contact matrix used, and finally section IV describes the treatment of case data (subsec. IV-A), the choice of initial conditions (subsec. IV-B]), the fitting procedure (subsec. IV-C) and the sensitivity analysis evaluated regarding pathogenicity and data period analysed (subsec. IV-D).

I. Model Equations

The model is an extended Susceptible, Exposed, Infected, and Recovered (*SEIR*) model that comprises susceptible (S), pre-symptomatic (E), asymptomatic (A), mild symptomatic (I), severe/hospitalized (H), recovered (R) and deceased (D) compartments. These compartments are duplicated to account for a second variant of SARS-CoV-2, and each of them is stratified into three age classes: young ($<$ 20 years), adults (20-59 years), and the elderly ($\geq$ 60 years). The "wild-type" classes represent all non-P.1 variantspresent, which do not seem to be variants of concern.

We assume that the second variant is capable of reinfecting individuals who have recovered from infection by the wild-type variant while the inverse is not possible; in the absence of data indicating this possibility, allowing reinfection by the wild-type variant on recovered of infection by P.1 would have negligible effect due to the small time window (3 months) considered in the present work. We also consider that a variant is not capable of reinfecting individuals recovered from the same lineage. Our model does not include vaccination due to low rates of vaccination in Brazil during the study time period.

To model the virus spread in the population, we assume that asymptomatic individuals have equal infectiousness compared to symptomatic ones, while pre-symptomatic individuals have reduced infectiousness represented by ω. To model behaviour, we assume that symptomatic individuals self-isolate themselves to some degree, reducing their contacts by ξ. Individuals with severe disease have greater isolation ξ_{sev} due to hospitalization. The daily contacts between each age class is represented by the matrix $\hat{C}$. The force of infection λ_k for each variant k is defined below:

$$\lambda_k = \beta_k \hat{C}[A_k + \omega E_k + (1-\xi)I_k + (1-\xi_{sev})H_k]$$

The complete system of equations is given by:

Completely Susceptible

$$\frac{dS}{dt} = -\lambda_1 \frac{S}{N} - \lambda_2 \frac{S}{N}$$

Wild-type variant

$$\frac{dE_1}{dt} = \lambda_1 \frac{S}{N} - \frac{E_1}{\gamma_1}$$

$$\frac{dA_1}{dt} = \frac{(1-\sigma_1)\alpha_1 E_1}{\gamma_1} - \frac{A_1}{\nu_{i,1}}$$

$$\frac{dI_1}{dt} = \frac{(1-\alpha_1)(1-\sigma_1)E_1}{\gamma_1} - \frac{I_1}{\nu_{i,1}}$$

$$\frac{dH_1}{dt} = \frac{\sigma_1 E_1}{\gamma_1} - \frac{H_1}{\nu_{s,1}}$$

$$\frac{dR_1}{dt} = \frac{A_1}{\nu_{i,1}} + \frac{I_1}{\nu_{i,1}} + \frac{(1-\mu_1)H_1}{\nu_{s,1}} - p_r \lambda_2 \frac{R_1}{N}$$

$$\frac{dD_1}{dt} = \frac{\mu_1 H_1}{\nu_{s,1}}$$

P.1 variant

$$\frac{dE_2}{dt} = \lambda_2 \frac{S}{N} - \frac{E_2}{\gamma_2} + p_r \lambda_2 \frac{R_1}{N}$$

$$\frac{dA_2}{dt} = \frac{(1-\sigma_2)\alpha_2 E_2}{\gamma_2} - \frac{A_2}{\nu_{i,2}}$$

$$\frac{dI_2}{dt} = \frac{(1-\alpha_2)(1-\sigma_2)E_2}{\gamma_2} - \frac{I_2}{\nu_{i,2}}$$

$$\frac{dH_2}{dt} = \frac{\sigma_2 E_2}{\gamma_2} - \frac{H_2}{\nu_{s,2}}$$

$$\frac{dR_2}{dt} = \frac{A_2}{\nu_{i,2}} + \frac{I_2}{\nu_{i,2}} + \frac{(1-\mu_2)H_2}{\nu_{s,2}}$$

$$\frac{dD_2}{dt} = \frac{\mu_2 H_2}{\nu_{s,2}}$$

Supplementary Equations

$$\frac{dC_1}{dt} = \chi(1-\alpha_1)\sigma_1 \frac{E_1}{\gamma_1}$$

$$\frac{dC_2}{dt} = \chi(1-\alpha_2)\sigma_2 \frac{E_2}{\gamma_2}$$

where C_1 ad C_2 are the cumulative hospitalization cases reported, and each variable of the system $(S,\ E_k, \ldots,\ C_k)$ is a vector containing each age class, e.g., $E_1 = (E_{1,y}, E_{1,a}, E_{1,e})^T$. The equations were numerically solved by the package developed by Soetaert et al (9).

II. Parameterization of the model

Table A.I presents parameters considered for the wild-type variant. The parameters for the P.1 variant are the same except for those considered in the model fitting.

Table A.I: Epidemiological parameters

Parameter	Description	Value	Source
γ	Average time in days between being infected and developing symptoms	5.8	Wei et al 2020 (10)
ν_i	Average time in days between being infectious and recovering for asymptomatic and mild cases	9	Cevik et al 2021 (11)
ν_s	Average time between being infectious and recovering/dying for severe cases	8.4	Average value based on SIVEP-Gripe data for SARI from Manaus, excluding outliers (12)
ξ_i	Reduction on the exposure of symptomatic cases (due to symptoms/quarantining)	0.1	Assumed
ξ_s	Reduction on the exposure of severe cases (due to hospitalization)	0.9	Assumed
ω	Relative infectiousness of pre-symptomatic individuals	1	Assumed
α	Proportion of asymptomatic cases	[0.67,0.44,0.31]	Young (13) Adult and Elderly (14)
σ	Proportion of symptomatic cases that require hospitalization	[0.001,0.012,0.089] *	Salje et al. (15)
μ	In-hospital mortality ratio	[0.417,0.188,0.754]	Portella et al. (16)
χ	Case report probability	1.0	Assumed

* The proportion is weighted by the age distribution of the population.

III. Contact Matrices

Our model includes three age group categories: young ($[0-19]y.o.$), adults ($[20-59]y.o.$) and elderly (greater than $60y.o.$). To model contacts between these groups we use estimated contact matrices computed by Prem et al (17), but since the original matrices use five-year age bins going up to 95+ years, we aggregate classes leading to a 3×3 matrix in the following way:

Let A, B be sets of indexes forming age groups (not necessarily of equal sizes), $x_{i,j}$ denoting contact between age groups i and j in the original matrix, d_i denoting population size of the age group i. The new contact matrix $\hat{C}$ is given by:

$$\hat{C}_{A^*,B^*} = \frac{\sum_{i \in A} \sum_{j \in B} d_i \, x_{i,j}}{\sum_{i \in A} d_i}$$

where A^*, B^* denotes a new indexation rule. Note that the contact matrices depend on local demographics and therefore must be computed for each place of study.

IV. Data Analysis Procedures

A. Nowcasting

Data used in parameters estimation were collected from the national public health system of severe acute respiratory illness (SARI) surveillance database, named *Sistema de Vigilância da Gripe - SIVEP-Gripe*. In this system, reporting of cases can be delayed for several reasons, including the notification system itself and confirmation of RT-PCR test results. The nowcasting procedure estimates, based on the past delay distribution, the number of cases that already occurred but were not yet reported. A window of 10 weeks is the acting window on the series, since delays greater than this are rare.

Nowcasting requires a pair of dates: (i) onset date of the event and (ii) report date of the event. The delay distribution is modeled as being best described as a Poisson distribution for days since the onset date to the report date. We considered *the first symptoms date* as the onset date. For the report date, we used the latest between *the test result date* and *the clinical classification date*. The nowcasting algorithm were developed by McGough et al 2020 (6) , and implemented in the *NobBS* (Nowcasting by Bayesian Smoothing) package in .

B. Initial Condition Estimation

The model requires appropriate mid-epidemic initial conditions in order to give relevant results. In the model, the number of new hospitalizations at a given time – h_{new}, is directly proportional to the number of exposed individuals at that time, therefore data was used to get an approximation of the number of exposed people. Also, to quantify the number of people belonging to the recovered class, prevalence was used.

We can estimate the appropriate initial conditions by finding an approximation for our model that relates more directly to the available data in each class. In the absence of the variant P.1, the model has four classes of infected compartments, namely $y = (E_1,\ A_1,\ I_1,\ H_1)^T$, and another three classes, represented by z, i.e., $z = (S,\ R_1,\ D_1)^T$.

To that effect, we can write the system as:

$$\dot{y} = F(y,z) - G(y,z),$$

$$\dot{z} = J(y,z)\ ,$$

where F comprises all entries of new Infected, coming from classes z, whilst G accounts for the transitions within infected classes and also recovery and death from the disease. Then, to find a good approximation for a small time window, we perform a linearization of our model around a point (y,z). Keeping z fixed, we get

$$\dot{y} = \left(\hat{F} - \hat{G}\right)y\ ,$$

where $\hat{F}$ and $\hat{G}$ are the linearized matrices arising from the functions F and G, respectively. The only entrance of new infected comes from the $\beta S\lambda/N$ terms in the $\dot{E}_1 = (\dot{E}_{1,y}, \dot{E}_{1,a}, \dot{E}_{1,e})^T$ equations (sub-indexes are y young, a adults and e elderly), then, the only non-zero elements of $\hat{F}$ are in its first 3 lines. Before proceeding, it's useful to define

$$\hat{b} = diag(S)\hat{C}$$

which allow us to write

$$\hat{F} = \frac{\beta}{N}\begin{bmatrix} \omega\,\hat{b} & \hat{b} & (1-\xi)\,\hat{b} & (1-\xi_{sev})\,\hat{b} \\ & & \mathbb{0}_{9,12} & \end{bmatrix}$$

$\hat{G}$ contains the terms of Exposed, E_1, the 3 possible forms of the disease considered in the model, that is A_1, I_1 and H_1, as the terms in its first 3 rows, whilst the remainder of its main diagonal contains terms of recovery and death. For simplicity, every constant (or vector for

the terms with σ) in $\hat{G}$ expression (A1) should be thought as diagonal matrices with its elements given by the constants (or vectors) and every $\mathbb{0}$ is a 3-dimensional square matrix where all entries are null.

$$\hat{G} = \begin{bmatrix} \gamma^{-1} & \mathbb{0} & \mathbb{0} & \mathbb{0} \\ -\alpha(1-\sigma)\gamma^{-1} & \nu_i^{-1} & \mathbb{0} & \mathbb{0} \\ -(1-\alpha)(1-\sigma)\gamma^{-1} & \mathbb{0} & \nu_i^{-1} & \mathbb{0} \\ -\sigma\gamma^{-1} & \mathbb{0} & \mathbb{0} & \nu_s^{-1} \end{bmatrix} \qquad \text{(A1)}$$

The linearization above implies that, for a small time interval, y has an exponential behavior and that the eigenvalues of $\hat{L} = \hat{F} - \hat{G}$ are related to the exponential growth rates. Therefore, a short time after the beginning of the epidemic, the largest eigenvalue should be the one to dominate. So the exponential growth rate of the wild-type variant – r, can be matched to the largest eigenvalue of $\hat{L}$ to obtain an estimate for β. The eigenvector associated with the largest eigenvalue gives the proportions of infected classes, which, together with the estimated number of exposed individuals – $E_1 = \frac{\gamma_1 h_{new}}{\sigma_1}$, results in an approximation for the number of people in the other infected classes.

Given a β, the largest eigenvalue of the linearization matrix is computed using the `eigs` function of the package *rARPACK* (18) and we find the β that gives r as the largest eigenvalue through bisection root finding. Finally, subtracting the number of recovered and infected from the total population gives the number of susceptible individuals.

C. Likelihood Estimation

Given the cumulative daily curves of hospitalization for wild-type variant, C_1, and P.1 variant, C_2, we can obtain the daily variation of each curve, namely ΔC_1^t and ΔC_2^t. Those curves are summed up to give the total number of weekly new cases:

$$\Delta C^\tau = \sum_{i=1}^{7} (\Delta C_1^{\tau-1+i} + \Delta C_2^{\tau-1+i}) \qquad \text{(A2)}$$

where τ is a discrete index given in weeks.

To calculate the frequency of P.1 in a given time period T, we use the proportion of new cases in this period from the wild-type and P.1 variant as follows:

$$P^{t'} = \frac{\sum_{i=1}^{T} \Delta C_2^{T-1+i}}{\sum_{i=1}^{T} \Delta C_1^{T-1+i} + \sum_{i=1}^{T} \Delta C_2^{T-1+i}} \qquad \text{(A3)}$$

where t' is a discrete index given in T periods.

The time period T depends on the dataset of genome sequences: it is daily in Faria et al (3) and monthly in Fiocruz website (7).

Using maximum likelihood, we fitted the model by estimating five parameters, namely, the relative transmissibility ($\Delta\beta$), the reinfection probability of P.1 (p), initial total prevalence ($\rho^0 = [R/N]_{t=0}$), initial fraction of cases that were caused by the new variant (P^0), and intrinsic growth rate of the wild-type variant (r). The parameter r incorporates effects related to contact rates for the wild-type variant, such as non-pharmacological interventions relaxation, elections, and others.

Number of hospitalization cases were assumed to follow a Poisson distribution, with expected value given by equation (A2). The recorded number of P.1 in genome samples was assumed to follow a binomial distribution with an expected value equal to the product of the total number of genome sequences sampled in each date and the proportion of P.1 cases (equation (A3)). The log-likelihood function for the model fitting was then:

$$\mathcal{L} = \sum_i \log \mathrm{Pois}\left(x^i \middle| \lambda = C^i\right) + \sum_j \log \mathrm{Bin}\left(y^j \middle| N = n^j, \theta\left(\pi^j\right) = P^j\right), \qquad \text{(A4)}$$

where *Pois* is a Poisson distribution with parameter λ, x^i is the number of recorded hospitalizations in week i, *Bin* is a Binomial distribution with parameters N (total number of trials) and π^j (probability of success at each trial), n^j is total number of sequences in clinical samples in week or day j, y^j is the number of P.1 sequences in each of these samples, and $\theta(.)$ is the *logit* function.

The model was then fitted by finding the values of the five above mentioned parameters that maximize the log-likelihood function (equation (A4)), using the function `mle2`, from the package *bbmle* (19).

To find starting values for the optimization performed by mle2 we calculated the log-likelihood function for a sample of one million combinations of parameters values within reasonable ranges. The sets of parameters values that provided the two higher values of the log-likelihood function were then used as starting values for the computational optimization.

The confidence intervals for the expected number of cases and frequency were estimated from 10000 parametric bootstrap samples assuming that the estimated parameters follow a multivariate normal distribution. The parameters of these multivariate distributions were the estimated values and estimated variance-covariance matrix of the parameters. For each sampled combination of parameters, the expected values were calculated and the confidence interval was estimated as the 2.5% and 95% quantiles of the distribution of bootstrapped expected values.

D. Sensitivity analysis

The model fitting assumed a constant Infection Hospitalization Rate (IHR, parameter σ) for each age group over time for both variants. An increase in IHR caused by P.1 would increase the number of hospitalizations even without any increase in transmissibility or reinfection. Because the pathogenicity of the variant P.1 is unknown, the model fitting was repeated assuming that the odds ratio of the IHR in each age class was twice for P.1 cases compared to wild-type variant cases (SA1). Moreover, as the collapse of Manaus health system hindered hospitalizations of new severe cases and may have affected case recording in surveillance databases, the model fitting was repeated considering only the period prior to the collapse (10 January 2021) (SA2).