Phenomics
https://doi.org/10.1007/s43657-022-00058-x

ARTICLE

Clinical Profiles at the Time of Diagnosis of SARS-CoV-2 Infection in Costa Rica During the Pre-vaccination Period Using a Machine Learning Approach

Jose Arturo Molina-Mora[1] · Alejandra González[2] · Sergio Jiménez-Morgan[3] · Estela Cordero-Laurent[2] · Hebleen Brenes[2] · Claudio Soto-Garita[2] · Jorge Sequeira-Soto[2] · Francisco Duarte-Martínez[2]

Received: 7 February 2022 / Revised: 22 April 2022 / Accepted: 27 April 2022

Abstract

The clinical manifestations of COVID-19, caused by the SARS-CoV-2, define a large spectrum of symptoms that are mainly dependent on the human host conditions. In Costa Rica, more than 169,000 cases and 2185 deaths were reported during the year 2020, the pre-vaccination period. To describe the clinical presentations at the time of diagnosis of SARS-CoV-2 infection in Costa Rica during the pre-vaccination period, we implemented a symptom-based clustering using machine learning to identify clusters or clinical profiles at the population level among 18,974 records of positive cases. Profiles were compared based on symptoms, risk factors, viral load, and genomic features of the SARS-CoV-2 sequence. A total of 18 symptoms at time of diagnosis of SARS-CoV-2 infection were reported with a frequency > 1%, and those were used to identify seven clinical profiles with a specific composition of clinical manifestations. In the comparison between clusters, a lower viral load was found for the asymptomatic group, while the risk factors and the SARS-CoV-2 genomic features were distributed among all the clusters. No other distribution patterns were found for age, sex, vital status, and hospitalization. In conclusion, during the pre-vaccination time in Costa Rica, the symptoms at the time of diagnosis of SARS-CoV-2 infection were described in clinical profiles. The host co-morbidities and the SARS-CoV-2 genotypes are not specific of a particular profile, rather they are present in all the groups, including asymptomatic cases. In addition, this information can be used for decision-making by the local healthcare institutions (first point of contact with health professionals, case definition, or infrastructure). In further analyses, these results will be compared against the profiles of cases during the vaccination period.

Keywords COVID-19 · Costa Rica · Machine learning · Diagnosis · SARS-CoV-2 · Clinical profiles

Alejandra González: Deceased.

✉ Jose Arturo Molina-Mora
jose.molinamora@ucr.ac.cr

Alejandra González
agonzalez@inciensa.sa.cr

Sergio Jiménez-Morgan
sergio.jimenezmorgan@ucr.ac.cr

Estela Cordero-Laurent
ecordero@inciensa.sa.cr

Hebleen Brenes
hbrenes@inciensa.sa.cr

Claudio Soto-Garita
csoto@inciensa.sa.cr

Jorge Sequeira-Soto
jsequeira@inciensa.sa.cr

Francisco Duarte-Martínez
fduarte@inciensa.sa.cr

[1] Centro de Investigación en Enfermedades Tropicales (CIET) and Facultad de Microbiología, Universidad de Costa Rica, San José 2060, Costa Rica

[2] Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud (INCIENSA), Tres Ríos 30301, Costa Rica

[3] Escuela de Medicina, Universidad de Costa Rica, San José 2060, Costa Rica

Published online: 07 June 2022

Introduction

The Coronavirus Disease 2019 (COVID-19) caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was declared as a pandemic and has impacted the public health systems around the world, even though a new hope was established with the beginning of the vaccination program at the end of 2020. In Costa Rica, more than 169,000 cases and 2185 deaths were reported during 2020, a pre-vaccination period (https://www.ministeriodesalud.go.cr/). In our previous work, we focused on the analysis of the genomic diversity of SARS-CoV-2 sequences from Costa Rica during 2020 (Molina-Mora et al. 2021; Molina-Mora 2022), and we now studied the symptoms at the time of diagnosis of SARS-CoV-2 infection as clinical profiles.

As an infectious disease, the spread and manifestations of COVID-19 depend on the agent (the SARS-CoV-2), the human host (comorbidities and genetic factors), and the environment (physical environmental conditions, social interactions, containment measures, etc.) (Tsui et al. 2020). In addition, COVID-19 diagnosis—apart from the place's epidemiological rates—requires both medical history and clinical manifestations, as well as radiologic and laboratory data (Mouliou et al. 2022a).

Apart from sole SARS-CoV-2 pre-/post-symptomatic or asymptomatic carriers, COVID-19 has shown various clinical manifestations, from paucisymptomatics (1 symptom only) and mild-to-moderate patients, to critical disease conditions (Nicastri et al. 2020; Mouliou and Gourgoulianis 2022). Symptomatic cases report a variety of symptoms, including fever, anosmia, cough, and diarrhea; more severe cases are reported with respiratory distress, sepsis, septic shock, and death (Huang et al. 2020). Due to the diversity of symptoms, human factors such as genetics and risk factors play a critical role in the outcome of the disease (LoPresti et al. 2020; Sironi et al. 2020; Toyoshima et al. 2020). These factors tend to be specific to the population, in which particular studies are required in each geographic location. In addition, many patients are evaluated only at the time of diagnosis due to the clinical presentation of a mild illness, in which the tracking of symptoms is lack or not possible later. This points out the need for defining clinical profiles at the initial stages of the COVID-19, for example at the time of diagnosis. Besides, although people can transmit the virus regardless of vaccination status and vaccination rates are highly influenced by the internet personal information, work and social life (Mouliou et al. 2021b), the vaccines have impacted the spread and the clinical manifestations of COVID-19 (Amit et al. 2021; Moghadas et al. 2021). In this situation, the vaccination period can be eventually contrasted with cases from the pre-vaccination pandemic time.

On the other hand, the diversity and mainly the amount of SARS-CoV-2-infected cases define a complex challenge in the step of data analysis to describe the clinical features in the populations. To overcome this situation, clustering or unsupervised machine learning approaches bring an opportunity to extract relevant information by identifying patterns, clusters, or profiles within large volumes of data. Although some machine learning or similar approaches have been implemented to investigate clinical symptoms, risk factors, and other parameters related to COVID-19 (Dixon et al. 2021; Han et al. 2020; Sudre et al. 2021; Tong et al. 2020; Kim et al. 2020; Fu et al. 2020; Alballa and Al-Turaiki 2021; Oshinubi et al. 2021, 2022; Quiroz-Juárez et al. 2021; Zoabi et al. 2021; Li et al. 2020), to our knowledge, none has been formally reported from Costa Rican cases.

Therefore, because of the relevance of describing local clinical profiles at the population level in the early stages of COVID-19 in a pre-vaccination pandemic period, and the use of strategies to deal with massive data, this work aimed to identify and describe clinical profiles at the time of diagnosis of SARS-CoV-2 infection in Costa Rica during 2020 with a symptom-based clustering approach using machine learning.

Materials and Methods

Data Source, Software, and Pre-processing

This is an observational retrospective study with SARS-CoV-2-infected cases from Costa Rica. Initially, 68,758 records of suspected patients were included. Data corresponded to all the registered cases in *Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud* (INCIENSA), the institution in charge of the epidemiological surveillance in Costa Rica) during the year 2020 (between March 6 and December 31, 2020).

All the different analyses for pre-processing, machine learning approaches, and visualization were performed with custom scripts in the RStudio software (Version 1.1.453, https://www.rstudio.com/) with the R software (Version 3.6.3, https://www.r-project.org/) in local servers of the Universidad de Costa Rica. The following packages were used during this implementation: "caret", "haven", "RColorBrewer", "ggfortify", "cluster", "plotrix", "ggpubr", and "randomcoloR" (details in https://cran.r-project.org/web/packages/).

For the pre-processing step, different filtering, cleaning, and re-arrangement strategies were applied to data, as follows. We only considered cases with positive results by real-time Reverse Transcription Polymerase Chain Reaction (rRT-PCR) test for SARS-CoV-2, without repeated tests (for patients with multiple tests, we only selected the first

record), completing 18,974 records. Each record was composed of 121 epidemiological and clinical (symptoms at the time of diagnosis and risk factors) data and the viral load by the C_t value in the rRT-PCR assay. For 160 cases, genomic information of the viral sequences of SARS-CoV-2 (clade and lineage, and the presence of the mutation spike-T1117I of the Costa Rican lineage B.1.1.389) was available from our previous work (Molina-Mora et al. 2021), which was included for the comparisons.

Clustering Analysis by a Machine Learning Approach

To identify major groups of SARS-CoV-2-infected cases based on the symptomatology at the time of diagnosis, a clustering analysis was completed with all the 18,974 records. Although there were 51 distinct symptoms among the patients, most of them were of very low frequency. Thus, we only included symptoms present in at least 1% of the patients, with a final selection of 18 symptoms to describe the manifestations at the population level. A small group of symptomatic patients with only "rare" or low-frequency symptoms was analyzed as non-symptomatic cases at this step.

Afterward, to define the groups based on the 18 frequent symptoms (frequency > 1%, see the complete list of symptoms in Fig. 2) of the 18,974 patients, a machine learning strategy was implemented using Hierarchical Clustering (HC). To select the best conditions for the clustering analysis, we followed three main steps. First, to define how different were the clinical manifestations among all patients, we assessed five different distance metrics (Euclidean, Binary, Maximum, Manhattan, and Minkowski). The optimal metric had to identify a separated group for the "asymptomatic cases". Second, the Elbow criterion was implemented to determine the expected number of major clusters, by plotting the explained variation as a function of the number of clusters (Shi et al. 2021). The number of clusters K was defined according to the elbow of the curve and, due to this is a heuristic approach, a tolerance of 1 was considered (i.e., number of clusters $= K \pm 1$). Finally, using the optimal distance metric and the expected number of clusters, the tree was cut using a single height value to define the clusters. Groups with at least 5% of the cases (949 out of the 18,974 patients) were labeled as major clusters, and the remaining small groups were included in a single "sink" cluster.

Clusters Comparison

After the definition of the major clusters, the groups were compared using demographic data (age, sex, localization, etc.), clinical information (symptoms, risk factors, vital status, hospitalization, C_t value, etc.), and SARS-CoV-2 genotypes (clades, lineages, and presence of the spike-T1117I mutation of the Costa Rican lineage B.1.1.389). To this end, representation of comparisons was done using heatmaps, barplots, and boxplots, with the subsequent statistical tests by ANOVA, Tukey test, Chi-square, and other tests as appropriate.

Results

Seven Major Clusters with Specific Symptoms Define the Clinical Profiles at the Time of Diagnosis of SARS-CoV-2 Infection in Costa Rica

To identify clinical profiles of SARS-CoV-2-infected cases based on 18 frequent symptoms (present in at least 1% of the patients) at the time of diagnosis, we developed a clustering strategy using machine learning with 18,974 records. After data pre-processing, five distance metrics were assessed within the HC algorithm. The selection of the best metric was based on the ability to separate all the asymptomatic cases in a single group, which was only achieved when a Binary distance was implemented (Fig. 1a). Other approaches using Euclidean or Manhattan distance (Fig. 1b, c), resulted in groups with symptomatic and asymptomatic cases into the major clusters. To define the number of expected clusters, the Elbow criterion suggested $k = 8 \pm 1$ as the optimal number of clusters to be generated using the whole data set (Fig. 1d).

Using the parameters for the optimal clustering (distance and number of expected clusters), 25 clusters were obtained, including very small groups. Seven clusters were composed by at least 5% of cases (represented by non-gray colors), and they were subsequently referred as major clusters. The percentage of cases for each major cluster was: C1 19.0%, C2 5.1%, C3, 14.1%, C4 11.1%, C5, 5.5%, C6 6.3%, and C7, 5.0%. The remaining small groups, which were defined when the clustering tree was cut, were joined into a sink group (dark gray). See Fig. 2 (top) for details of the clusters, and Supplementary file for details of the size for all the clusters. In Fig. 2, the red cluster corresponded to the group with all the asymptomatic cases. As found in the heatmaps for all the patients (Fig. 2) and the total frequency (Fig. 3, left), the composition is dependent on the symptoms, as expected. See below for more details.

As shown in Table 1, major clusters are composed of between 953 and 3,613 patients (also see Fig. 4a). The 3440 cases without any of the 18 main symptoms were found in cluster C1. The small fraction of 173 symptomatic cases in the C1 is the patients with "rare" or low-frequency symptoms (not included in the 18 used for the clustering), as expected. No other patterns regarding age, vital status, sex, nor hospitalization conditions were recognized, and

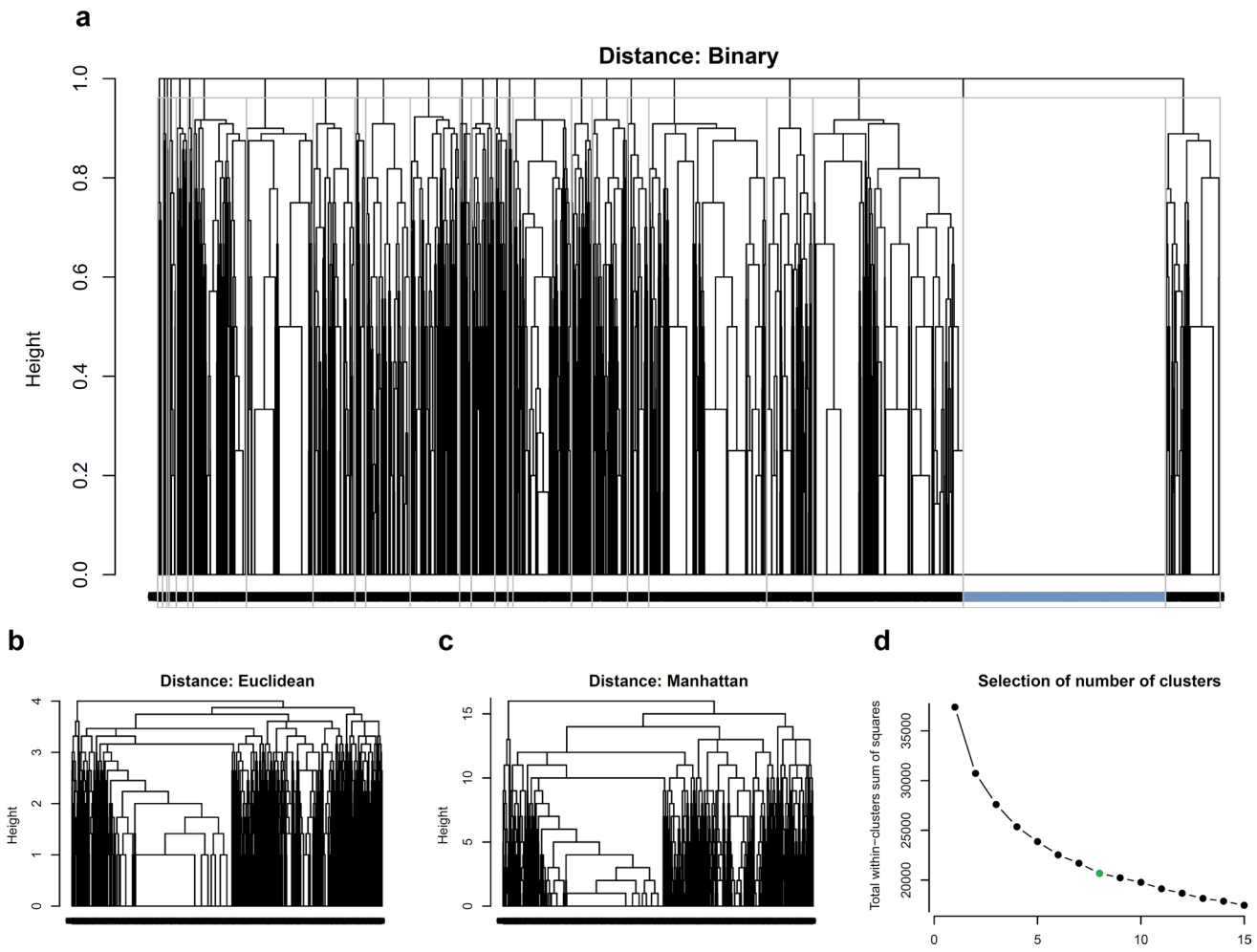

Fig. 1 Parameters of the clustering using machine learning to identify clinical profiles of SARS-CoV-2-infected cases based on symptoms at the time of diagnosis. Using clinical data of 18,974 patients, different clustering analyses were run with different distance metrics, including Binary (**a**), Euclidean (**b**), and Manhattan (**c**). Only the Binary distance was able to cluster the asymptomatic cases in a single group, as expected (blue group). In the analysis using the Elbow criterion (**d**), the plot of variation identified the $k=8$ (green) as the number of expected clusters

these parameters were distributed similarly among clusters (Table 1, Fig. 4b and Supplementary Fig. S1).

Analysis of the co-presence of symptoms among the patients (Fig. 2, columns), several symptoms were clustered (rows, left side). For example, there is a cluster of general symptoms (Fig. 2, top left), digestive conditions (middle left), or more respiratory symptoms (down left).

In the comparison between symptoms (Fig. 3, left), each cluster has a specific clinical profile. Cluster C1 is the group of all the asymptomatic cases. The C2 is characterized mainly by the presence of cough and rarely other symptoms. In contrast, C3 and C4 include cough and another main symptom (fever and headache, respectively). C5 is mainly composed of four symptoms, including arthralgia as the header. The conditions of anosmia and dysgeusia are the major components of the C6 and C7 clusters, with an inverted pattern of frequency.

Risk Factors and Diverse SARS-CoV-2 Genomes are Distributed Among All the Clinical Profiles, and Viral Load Inferred from C_t Values was Lower for Asymptomatic Cases

Concerning the description of the risk factors among the clusters (Fig. 3 right), all the conditions are present in all the groups without specific patterns, including the C1 for asymptomatic patients and the sink. The conditions with higher frequency are high blood pressure (HBP), asthma, and diabetes among all the profiles. Interestingly, asthma

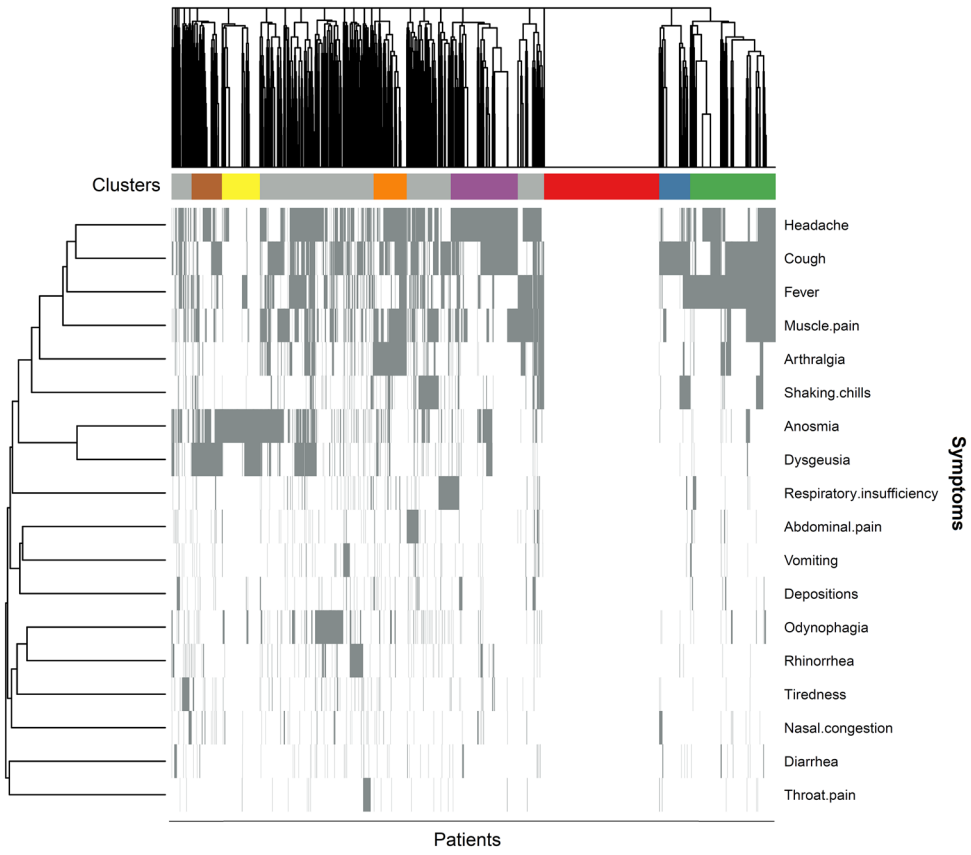

Fig. 2 Seven major clinical profiles of SARS-CoV-2-infected cases were identified by a clustering approach using symptom information at the time of diagnosis. Seven major clusters (colors) and a sink group (dark gray) were defined, including a well-identified group for all the asymptomatic cases. Some symptoms co-occurred among patients (left dendrogram). In the heatmap, the presence or the absence of the symptom was represented by a light gray or white color, respectively

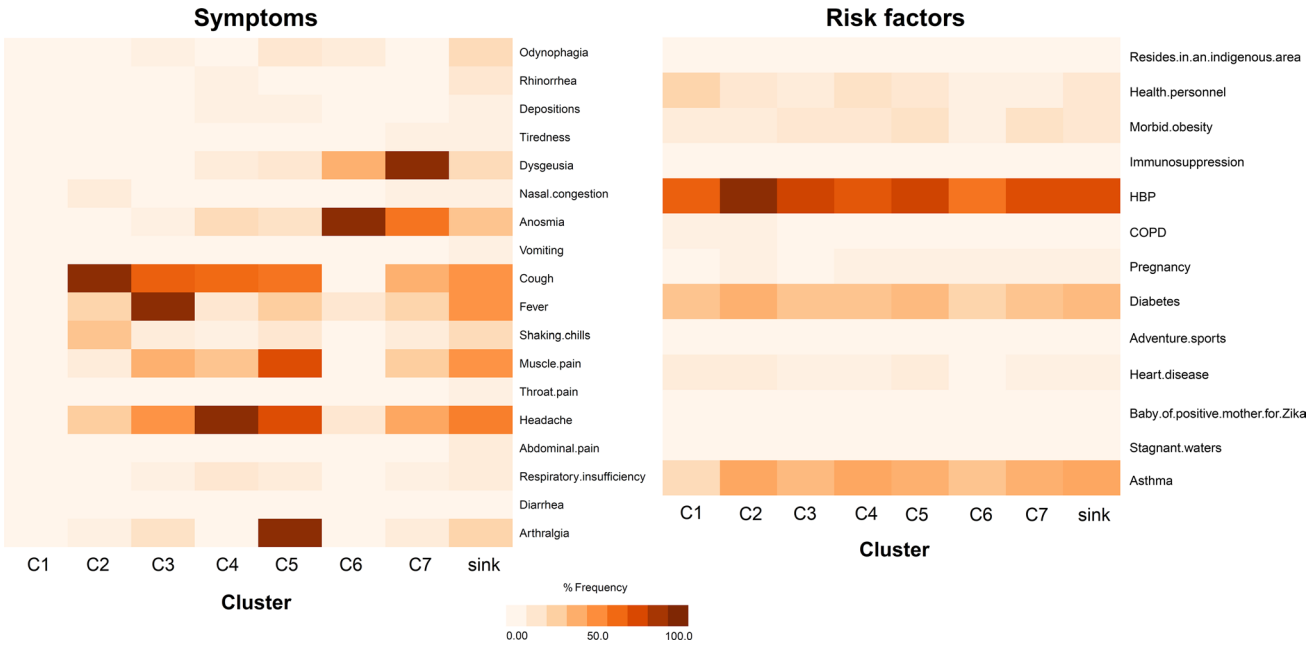

Fig. 3 Frequency patterns of symptoms and risk factors of patients among the clusters of the clinical profile. Each cluster is composed of specific and predominant symptoms (left). The risk factors are distributed among all the clusters without any enriched pattern, including the asymptomatic and sink groups

Table 1 Composition of clusters by epidemiological and genomic data

Groups	Clusters							
	C1	C2	C3	C4	C5	C6	C7	Sink
Total patients	3613	974	2683	2106	1042	1190	953	6409
Sex								
Female	1776	460	1133	1119	510	616	506	3217
Male	1819	511	1541	979	529	572	444	3175
ND	18	3	9	8	3	2	3	17
Symptoms								
Yes	173	974	2683	2106	1042	1190	953	6409
No	3440	0	0	0	0	0	0	0
Hospitalized								
Yes	52	19	42	34	11	13	15	105
No	1260	304	716	529	220	353	243	1553
ND	2301	651	1925	1543	811	824	695	4751
Alive (vital status)								
Yes	2840	792	2225	1772	896	973	830	5429
No	9	0	7	0	0	1	1	7
ND	764	182	451	334	146	216	122	973
Number of distinct GISAID clades[a]	3	4	3	3	3	1	2	4
Number of PANGOLIN lineages[a]	7	6	10	4	5	1	4	8
Presence of the mutation spike-T1117I[a]								
Yes	11	5	7	7	3	2	1	14
No	34	8	30	12	4	0	3	19

ND no data

[a]Based on 160 genomes

was found in a less frequency for the asymptomatic group, and HBP has a higher frequency in patients of cluster C2.

About the expected viral load (Fig. 4c), interestingly the C_t values for cluster C1 of asymptomatic cases were higher in comparison to all the other clusters ($p < 0.05$). See statistical details in the Supplementary Material.

On the other hand, using 160 cases in which the SARS-CoV-2 genome was sequenced, it was possible to infer that the SARS-CoV-2 clades and lineages were not associated with specific symptoms nor clinical profiles, and they are distributed among all the clusters (Fig. 4d, e). This also applies to the Costa Rican lineage B.1.1.389 (orange in the barplots of Fig. 4e), which carries the mutation spike-T1117I and was the most common detected lineage during 2020 in the country (Fig. 4f), which is not specific to a particular profile.

Discussion

To describe the clinical presentations at the time of diagnosis of SARS-CoV-2 infection in Costa Rica during the pre-vaccination period, we implemented a machine learning strategy to identify clusters or clinical profiles at the population level among 18,974 records of positive cases. Seven clinical profiles were identified with a specific composition of clinical manifestations and some patterns related to viral load (lower for the asymptomatic group), while the risk factors and the SARS-CoV-2 genomic features were distributed among all the clusters. This work was an observational study with random samples at random points of time for suspected cases in which most of them were on an ongoing disease and reported symptomatology.

The clinical manifestations of COVID-19 define a large spectrum of symptoms at the population level, as found in other studies (Kim et al. 2020; Fu et al. 2020; Sudre et al. 2021). Estimates of the features and proportion of the distinct clinical manifestations of COVID-19, including asymptomatic cases, are vital parameters for modeling studies (Byambasuren et al. 2020). In addition, early identification of symptoms is important for successful diagnosis, medical management, and treatment selection (Kostopoulou et al. 2015). This is a key point for health professionals that are in charge of gathering symptoms information when testing patients (the time of diagnosis during the first point of contact), to be able to differentiate between the most and least prevalent clinical presentation of COVID-19 in a specific community. In this regard, we were motivated to conduct this study with symptoms in SARS-CoV-2-infected cases from Costa Rica during 2020 (the pre-vaccination period).

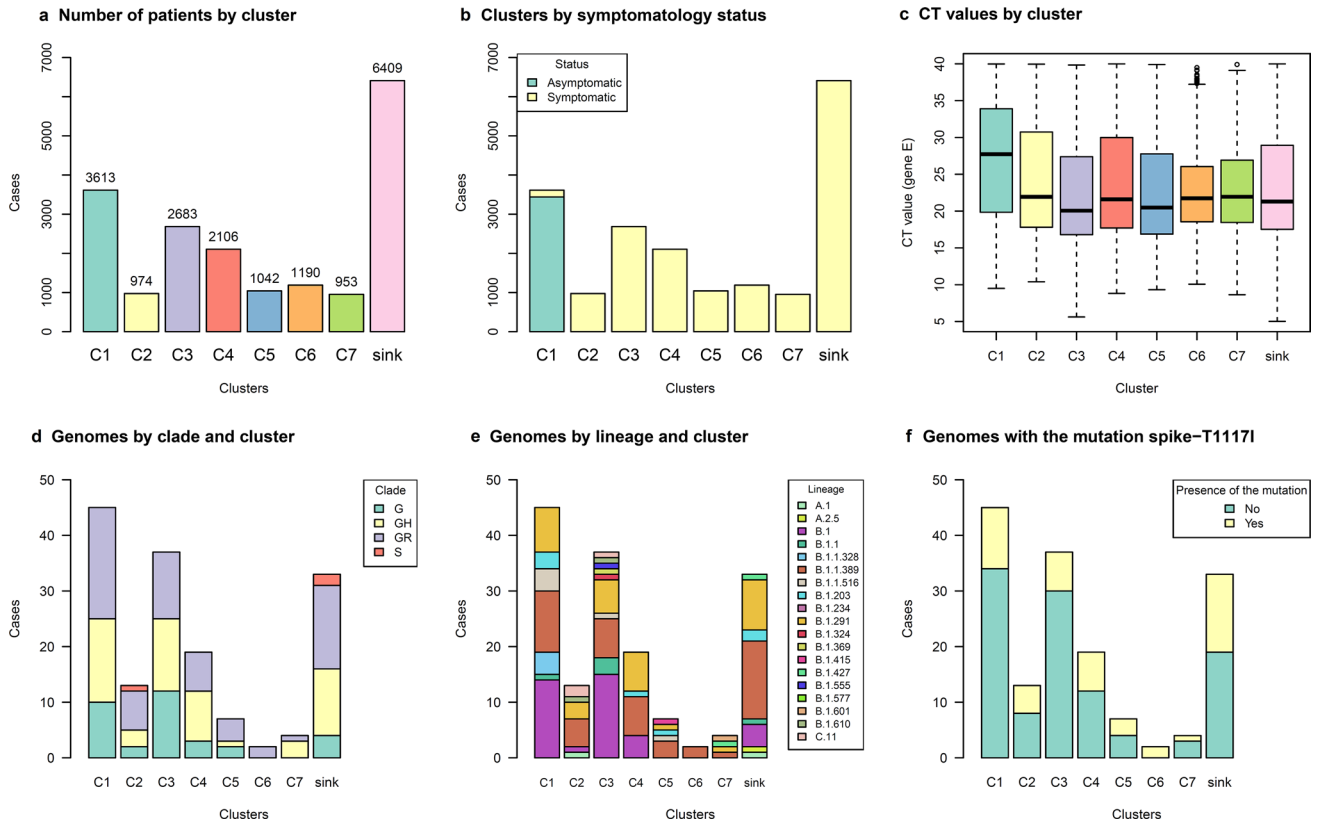

Fig. 4 Distribution of demographic, clinical, and SARS-CoV-2 genomic information of cases of SARS-CoV-2-infected cases among seven major clusters. Major clusters are composed of 953–3613 patients. All the asymptomatic cases are found in the same cluster C1. Interestingly, the viral load (inferred from the C_t value) is lower in this group. Different SARS-CoV-2 genomes (lineages, clades, and the presence of the mutation T1117I in the spike) were distributed among all the clinical profiles

At the time of diagnosis, 18 symptoms were found to be present in at least 1% of the SARS-CoV-2-infected cases from Costa Rica, including non-specific symptoms (fever, headache, etc.), as well as respiratory and gastrointestinal manifestations. Using a machine learning approach, seven major clusters or clinical profiles were found with those symptoms to describe the manifestations at the population level. The clusters showed the expected heterogeneity in the clinical presentation among SARS-CoV-2-infected cases from Costa Rica, just as it has been observed worldwide according to hundreds of case reports (Dixon et al. 2021; Han et al. 2020; Sudre et al. 2021; Tong et al. 2020; Kim et al. 2020; Fu et al. 2020). Besides, six main symptoms are defining the clinical profiles (Fig. 3) and that must be taken into higher consideration at the moment of filling a patient's chart: cough, fever, headache, arthralgia, anosmia, and dysgeusia. Congruently, most of these manifestations are included in the limited number of symptoms that are known to be associated with infectious diseases (Jeon et al. 2020). In addition, the general description of the clinical manifestations can be used as part of the "case definition of COVID-19" given by the local and international epidemiological surveillance systems (World Health Organization 2021).

A multivariable logistic regression and exploratory factor analysis by Dixon et al. (2021) determined five symptom clusters among which ageusia, anosmia, and fever tend to be highly associated with SARS-CoV-2 infection, which resembles our findings in cluster C6. This also supports other findings in a meta-analysis in which up to 52.73% and 43.93% of SARS-CoV-2-infected cases presented olfactory and gustatory dysfunction, respectively (Tong et al. 2020), also found in cluster C7. In a second cluster, Dixon et al. (2021) reported shortness of breath, cough, and chest pain, but only the cough had a high frequency in our data (cluster C2) without being associated with those other two symptoms. Interestingly, another study showed that a diversity of respiratory symptoms were found as a significant predictor for test positivity for the diagnosis of SARS-CoV-2 infection (Kotsiou et al. 2021).

A third cluster was composed of fatigue, muscle ache, and headache. Of those symptoms, we only found headache as the main symptom in cluster C4. Finally, the last two clusters reported were represented by vomiting and diarrhea,

and a runny nose with a sore throat (Dixon et al. 2021). None of those two clusters coincides with our findings. As Fig. 3 shows, even if digestive symptoms are present among Costa Rican SARS-CoV-2-infected cases from C1 to C7, their frequency is very low. Nonetheless, this should not be neglected as it has been reported that some individuals present digestive symptoms alone, which is of clinical relevance as those patients may last longer achieving viral clearance compared to those with associated respiratory symptoms (Han et al. 2020).

In another work, a similar approach with machine learning techniques for the study of COVID-19 symptoms, six temporal profiles were identified after self-reported data were used (Sudre et al. 2021). To make a better comparison, day 0 symptoms were contrasted with our findings. Interestingly, dysgeusia was not included as the main symptom in their study, even though it was the most prevalent one in our cluster C7. Cough and fever were found to be associated with the second cluster reported by Sudre et al. (2021) as well as in profile C3 in our study. Headaches were distributed among all the clusters in both studies.

About risk factors, three chronic diseases were found among Costa Rican patients in all of the seven clusters. From most to least prevalent, the most significant conditions were high blood pressure, diabetes, and asthma. Interestingly, this finding is highly consistent with a meta-analysis by Yang et al. (2020), who reported that the most prevalent comorbidities among SARS-CoV-2 patients were hypertension (21.1%), diabetes (9.7%), cardiovascular disease (8.4%), and respiratory system disease (1.5%). Another study, which was based on environmental and health-related predictors for SARS-CoV-2 infection, revealed a vulnerability to COVID-19 in cases with previous pneumonia (Mouliou et al. 2021a), although this risk factor was not studied in our work. Jointly, it is clinically relevant to take these comorbidities into account when performing a screening among COVID-19 tests. However, we identified no reliance on the co-morbidities and the clinical profiles for SARS-CoV-2-infected cases. This result is in line with a meta-analysis that reported that up to 90% of clinical and demographic variables showed inconsistent associations with COVID-19 outcomes (Jeon et al. 2020).

Despite consulting several databases, no other works using machine learning were found using symptoms, risk factors nor SARS-CoV-2 genomic data of SARS-CoV-2-infected cases at the same time, and none using the initial clinical profile at the time of diagnosis. Machine learning techniques prove to be a very useful approach to study the variety of COVID-19 symptoms when large sets of data are available. The heterogeneity of this disease's clinical presentation is reduced using this technique, thus it may help clinicians heighten vigilance of some specific symptoms over others.

On the other hand, the cluster of asymptomatic cases (C1) represents 18% of the total positive cases. This percentage is in line with other analyses in which the asymptomatic cases vary between 15 and 30% (Centers for Disease Control and Prevention US 2021; Byambasuren et al. 2020), although other studies found higher frequencies (Byambasuren et al. 2020; Lee et al. 2020). This variation can be explained by multiple factors, such as the: (i) definition of asymptomatic, which depends on a specific moment (at time of diagnosis in our case) but can eventually change during the infection with distinct symptom onset into pre-/post-symptomatic cases (Mouliou and Gourgoulianis 2022); in fact, these conditions have questioned the real existence of asymptomatic cases, as discussed in a recent study (Mouliou and Gourgoulianis 2022); (ii) diagnosis tests identify SARS-CoV-2 carriers and not necessarily COVID-19 patients (Mouliou et al. 2022b), and (iii) the possible existence of preexistent immunity by previous infection, that can affect the clinical outcome during reinfections or coinfections, although associations in the possible reduction of symptomatology are still being monitored (Mouliou and Gourgoulianis 2022; Molina-Mora et al. 2022).

The comparison of expected viral load between symptomatic and asymptomatic cases, using the C_t value, has been also reported as very variable (Tutuncu et al. 2021; Trunfio et al. 2021). Similar to our findings in which the symptomatic groups had lower C_t values, another study reported that higher viral load was associated with more signs and symptoms at diagnosis and a more frequent pattern of respiratory and systemic complaints (Trunfio et al. 2021). However, no associations between viral load and symptoms state have been also suggested in other works (Tutuncu et al. 2021; Lee et al. 2020). The situation of very diverse patterns of C_t values and clinical outcome is a drawback that can be explained by individual factors and technical issues. For example for a specific case (individual factors), as discussed before, some asymptomatic cases could be related to genetics, risk factors, or preexistent immunity by previous/concomitant infections (with possible effects on the viral replication or viral shedding and finally on symptoms onset or transmission) (Mouliou and Gourgoulianis 2022). In the case of sample processing (technical issues), the results can be affected by the technology, sample quality, and the time of sampling after infection (Buchan et al. 2020), as well as the general performance of the rRT-PCR test which is not errors-free and false positive and false negative results can be generated (Mouliou and Gourgoulianis 2021). Therefore, this complex scenario implies that there is no consensus between the initial viral load and the clinical manifestations of COVID-19 (Trunfio et al. 2021; Byambasuren et al. 2020).

Regarding the SARS-CoV-2 genotypes, our reports of the independence of the clinical presentation of

COVID-19 and the genomic determinants of the SARS-CoV-2 sequence are in line with others studies (Hodcroft et al. 2020; Grubaugh et al. 2020; van Dorp et al. 2020). For each cluster, a diversity of clades and lineages were identified, including independence of the presence or absence of the mutation T1117I from the Costa Rican lineage B.1.1.389 (Molina-Mora et al. 2021). This situation reminds us that the clinical profiles depend on the viral agent and human host conditions. The human genetic, comorbidities and risk conditions have been described as the predominant factor in the clinical outcome of the COVID-19, as found in several studies (LoPresti et al. 2020; Sironi et al. 2020; Toyoshima et al. 2020; Molina-Mora et al. 2021).

Furthermore, owing to the distribution of SARS-CoV-2 genotypes among all the clusters, our results suggest that genomic features of the virus are not associated with specific changes in the clinical presentation, as has been reported recently, including relevant variants (Nakamichi et al. 2021; Graham et al. 2021). The lack of change in symptoms for different SARS-CoV-2 genotypes also indicates that existing testing and surveillance infrastructure do not need to change specifically for these versions of the SARS-CoV-2 genome (Graham et al. 2021).

Our analyses presented some limitations that must be taken into account in the interpretation of results: (1) classification of positive cases of COVID-19 was based on the positivity of a rRT-PCR for nasopharyngeal samples, i.e., we depended on the performance of the test and sample quality; (2) C_t values were obtained by distinct RT-PCR test kits (not performed in the same lab and protocols), thus C_t value comparisons can be affected by these differences; (3) records were retrieved from a local database (with predefined symptoms) with some missing information, mainly for SARS-CoV-2 genomic data or other potential symptoms (i.e., dyspnea, sputum production, neck pain, shiver); (4) this study was based on symptoms present in at least 1% of the patients and rare or low frequency symptoms were not included for the clustering analysis (see "Materials and Methods"); (5) cases corresponded to random samples at random points of time which were considered as a same group, without consideration of a particular symptomatology for reinfections or coinfections cases; and (6) data for social behavior or genetic factors of the host were not considered in this study.

Finally, due to vaccination started massively in January 2021 in Costa Rica (first doses were applied at the end of December 2020), we consider that this study represents a special work to give the panorama of COVID-19 in pre-vaccination time (2020). In future work, we hope to assess the vaccination status and how this event has influenced the clinical profiles of SARS-CoV-2-infected cases during 2021.

Conclusions

The identification of seven clinical profiles at the time of diagnosis of SARS-CoV-2 infection was achieved using a clustering approach. In general, at the population level, there were 18 symptoms reported in at least 1% of the SARS-CoV-2-infected cases from Costa Rica, although six clinical manifestations were predominant. A specific symptom frequency was revealed for each cluster or clinical profile. In the comparison between clusters, a higher viral load inferred from the C_t values was found for the asymptomatic group, while the risk factors and the SARS-CoV-2 genomic features were distributed among all the clusters. Therefore, the host co-morbidities and the SARS-CoV-2 genotypes are not specific of a particular profile, rather they are present in all the groups, including asymptomatic cases. No other distribution patterns were found for age, sex, vital status, and hospitalization.

Jointly, these results describe the clinical manifestations at the time of diagnosis of SARS-CoV-2 infection in Costa Rican patients during the pre-vaccination time of the pandemic, as well as they can be used for decision making by the local healthcare institutions (first point of contact with health professionals, case definition, or infrastructure). In further analyses, these clinical patterns will be compared against cases during the vaccination period.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43657-022-00058-x.

Acknowledgements We thank clinicians, microbiologists, and other personnel of the public (Caja Costarricense de Seguro Social CCSS) and private clinical laboratories for the samples of confirmed cases of COVID-19. We also thank members of CIET-Universidad de Costa Rica and INCIENSA for their logistic and financial support in the activities associated with the project.

Authors' Contributions JMM, HB, CSG, and FDM participated in the conception and design of the study. JSS was responsible for data acquisition from INCIENSA database. JMM and AG were involved in data pre-processing. JMM implemented and standardized all the machine learning pipelines. JMM, SJM, ECL, HB, CSG, JSS, and FDM were involved in the interpretation of results. JMM drafted the manuscript. All authors reviewed and approved the final manuscript.

Funding This work was funded by Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud (INCIENSA, DG-of-2020-174) and Vicerrectoría de Investigación—Universidad de Costa Rica, with the Project "C0196 Protocolo bioinformático y de inteligencia artificial para el apoyo de la vigilancia epidemiológica basada en laboratorio del virus SARS-CoV-2 mediante la identificación de patrones genómicos y clínico-demográficos en Costa Rica (2020–2022)".

Material Availability Processed data is found in the Supplementary material.

Declarations

Conflicts of Interest The authors declare that there is no conflict of interest.

Ethical Approval and Consent to Participate This study was approved by INCIENSA (INCIENSA-DG-of-2020-174) and the scientific committee of CIET-UCR (No. 242-2020). Data were collected for epidemiological surveillance according to the Costa Rican regulation Law No. 8270 (May 17th, 2002), in which no additional consent was required for retrospective studies of archived and anonymized samples.

Consent to Publish The authors provide the permission to publish this work.

References

Alballa N, Al-Turaiki I (2021) Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review. Inform Med Unlocked. https://doi.org/10.1016/j.imu.2021.100564

Amit S, Beni SA, Biber A, Grinberg A, Leshem E, Regev-Yochay G (2021) Postvaccination COVID-19 among healthcare workers, Israel. Emerg Infect Dis 27(4):1220–1222. https://doi.org/10.3201/eid2704.210016

Buchan BW, Hoff JS, Gmehlin CG, Perez A, Faron ML, Silvia Munoz-Price L, Ledeboer NA (2020) Distribution of SARS-CoV-2 PCR cycle threshold values provide practical insight into overall and target-specific sensitivity among symptomatic patients. Am J Clin Pathol 154(4):479–485. https://doi.org/10.1093/AJCP/AQAA133

Byambasuren O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P (2020) Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. J Assoc Med Microbiol Infect Dis Can 5(4):223–234. https://doi.org/10.3138/jammi-2020-0030

Centers for Disease Control and Prevention US (2021) COVID-19 pandemic planning scenarios | CDC. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html. Accessed 3 June 2021

Dixon BE, Wools-Kaloustian KK, Fadel WF, Duszynski TJ, Yiannoutsos C, Halverson PK, Menachemi N (2021) Symptoms and symptom clusters associated with SARS-CoV-2 infection in community-based populations: results from a statewide epidemiological study. PLoS ONE 16(3 March):1–13. https://doi.org/10.1371/journal.pone.0241875

Fu L, Wang B, Yuan T, Chen X, Ao Y, Fitzpatrick T, Li P et al (2020) Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis. J Infect 80(6):656–665. https://doi.org/10.1016/j.jinf.2020.03.041

Graham MS, Sudre CH, May A, Antonelli M, Murray B, Varsavsky T, Kläser K et al (2021) Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. Lancet Public Health 6(5):e335–e345. https://doi.org/10.1016/s2468-2667(21)00055-4

Grubaugh ND, Hanage WP, Rasmussen AL (2020) Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell 182(4):794–795. https://doi.org/10.1016/j.cell.2020.06.040

Han C, Duan C, Zhang S, Spiegel B, Shi H, Wang W, Zhang L et al (2020) Digestive symptoms in COVID-19 patients with mild disease severity: Clinical Presentation, Stool Viral RNA Testing, and Outcomes. Am J Gastroenterol 115(6):916–923. https://doi.org/10.14309/ajg.0000000000000664

Hodcroft EB, Zuber M, Nadeau S, Comas I, Candelas FG, SeqCOVID-SPAIN Consortium, Stadler T, Neher RA (2020) Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. MedRxiv 2020 (October). https://doi.org/10.1101/2020.10.25.20219063.

Huang C, Wang Y, Li X, Ren L, Zhao J, Yi Hu, Zhang Li et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Jeon J, Baruah G, Sarabadani S, Palanica A (2020) Identification of risk factors and symptoms of COVID-19: analysis of biomedical literature and social media data. J Med Internet Res 22(10):e20509. https://doi.org/10.2196/20509

Kim GU, Kim MJ, Ra SH, Lee J, Bae S, Jung J, Kim SH (2020) Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19. Clin Microbiol Infect 26(7):948.e1-948.e3. https://doi.org/10.1016/j.cmi.2020.04.040

Kostopoulou O, Rosen A, Round T, Wright E, Douiri A, Delaney B (2015) Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. Br J Gen Pract 65(630):e49-54. https://doi.org/10.3399/bjgp15X683161

Kotsiou OS, Pantazopoulos I, Papagiannis D, Fradelos EC, Kanellopoulos N, Siachpazidou D, Kirgou P et al (2021) Repeated antigen-based rapid diagnostic testing for estimating the coronavirus disease 2019 prevalence from the perspective of the workers' vulnerability before and during the lockdown. Int J Environ Res Public Health 18(4):1–12. https://doi.org/10.3390/ijerph18041638

Lee S, Kim T, Lee E, Lee C, Kim H, Rhee H, Park SY et al (2020) Clinical course and molecular viral shedding among asymptomatic and symptomatic patients with SARS-CoV-2 infection in a community treatment center in the Republic of Korea. JAMA Intern Med 180(11):1447–1452. https://doi.org/10.1001/jamainternmed.2020.3862

Li WT, Ma J, Shende N, Castaneda G, Chakladar J, Tsai JC, Apostol L et al (2020) Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med Inform Decis Mak 20(1):1–13. https://doi.org/10.1186/s12911-020-01266-z

LoPresti M, Beck DB, Duggal P, Cummings DAT, Solomon BD (2020) The role of host genetic factors in coronavirus susceptibility: review of animal and systematic review of human literature. Am J Hum Genet. https://doi.org/10.1016/j.ajhg.2020.08.007

Moghadas SM, Vilches TN, Zhang K, Wells CR, Shoukat A, Singer BH, Meyers LA et al (2021) The impact of vaccination on coronavirus disease 2019 (COVID-19) outbreaks in the United States. Clin Infect Dis 73(12):2257–2264. https://doi.org/10.1093/cid/ciab079

Molina-Mora JA (2022) Insights into the mutation T1117I in the spike and the lineage B.1.1.389 of SARS-CoV-2 circulating in Costa Rica. Gene Rep 27(January):1–9. https://doi.org/10.1016/J.GENREP.2022.101554

Molina-Mora JA, Cordero-Laurent E, Godínez A, Calderón-Osorno M, Brenes H, Soto-Garita C, Pérez-Corrales C et al (2021) SARS-CoV-2 genomic surveillance in Costa Rica: evidence of a divergent population and an increased detection of a spike T1117I mutation. Infect Genet Evol 92(August):104872. https://doi.org/10.1016/j.meegid.2021.104872

Molina-Mora JA, Cordero-Laurent E, Chacón-Ramírez E, Duarte-Martínez F (2022) Metagenomic pipeline for identifying co-infections among distinct SARS-CoV-2 variants of concern: study cases from alpha to omicron. Res Sq (Pre-Print), February. https://doi.org/10.21203/RS.3.RS-1389767/V1

Mouliou DS, Gourgoulianis KI (2021) False-positive and false-negative COVID-19 cases: respiratory prevention and management strategies, vaccination, and further perspectives. Expert Rev

Respir Med 15(8):993–1002. https://doi.org/10.1080/17476348.2021.1917389
Mouliou DS, Gourgoulianis KI (2022) COVID-19 'asymptomatic' patients: an old wives' tale. Expert Rev Respir Med 16(4):399–407. https://doi.org/10.1080/17476348.2022.2030224
Mouliou DS, Kotsiou OS, Gourgoulianis KI (2021a) Estimates of COVID-19 risk factors among social strata and predictors for a vulnerability to the infection. Int J Environ Res Public Health 18(16):8701. https://doi.org/10.3390/IJERPH18168701
Mouliou DS, Pantazopoulos I, Gourgoulianis KI (2021b) Social response to the vaccine against COVID-19: the underrated power of influence. J Pers Med 12(1):15. https://doi.org/10.3390/JPM12010015
Mouliou DS, Pantazopoulos I, Gourgoulianis K (2022a) COVID-19 smart diagnosis in the emergency department: all-in in practice. Expert Rev Respir Med 16(3):263-272. https://doi.org/10.1080/17476348.2022.2049760
Mouliou DS, Pantazopoulos I, Gourgoulianis K (2022b) COVID-19 diagnosis in the emergency department: seeing the tree but losing the forest. Emerg Med J 212219. https://doi.org/10.1136/EMERMED-2021-212219
Nakamichi K, Shen JZ, Lee CS, Lee A, Roberts EA, Simonson PD, Roychoudhury P et al (2021) Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. Sci Rep 11(1):4802. https://doi.org/10.1038/s41598-021-82850-9
Nicastri E, D'Abramo A, Faggioni G, De Santis R, Mariano A, Lepore L, Molinari F et al (2020) Coronavirus disease (COVID-19) in a paucisymptomatic patient: epidemiological and clinical challenge in settings with limited community transmission, Italy, February 2020. Eurosurveillance 25(11):2000230. https://doi.org/10.2807/1560-7917.ES.2020.25.11.2000230
Oshinubi K, Rachdi M, Demongeot J (2021) Analysis of reproduction number R0 of COVID-19 using current health expenditure as gross domestic product percentage (CHE/GDP) across countries. Healthcare 9(10):1247. https://doi.org/10.3390/HEALTHCARE9101247
Oshinubi K, Rachdi M, Demongeot J (2022) Modeling of COVID-19 pandemic vis-à-vis some socio-economic factors. Front Appl Math Stat 7(January):78. https://doi.org/10.3389/FAMS.2021.786983/BIBTEX
Quiroz-Juárez MA, Torres-Gómez A, Hoyo-Ulloa I, de León-Montiel RDJ, U'Ren AB (2021) Identification of high-risk COVID-19 patients using machine learning. PLoS ONE 16(9):e0257234. https://doi.org/10.1371/JOURNAL.PONE.0257234
Shi C, Wei B, Wei S, Wang W, Liu H, Liu J (2021) A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP J Wirel Commun Netw 2021(1):1–16. https://doi.org/10.1186/s13638-021-01910-w
Sironi M, Hasnain SE, Rosenthal B, Phan T, Luciani F, Shaw MA, Anice Sallum M, Mirhashemi ME, Morand S, González-Candelas F (2020) SARS-CoV-2 and COVID-19: a genetic, epidemiological, and evolutionary perspective. Infect Genet Evol 84(May):104384. https://doi.org/10.1016/j.meegid.2020.104384
Sudre CH, Lee KA, Lochlainn MN, Varsavsky T, Murray B, Graham MS, Menni C et al (2021) Symptom clusters in COVID-19: a potential clinical prediction tool from the COVID symptom study app. Sci Adv 7(12):1–7. https://doi.org/10.1126/sciadv.abd4177
Tong JY, Wong A, Zhu D, Fastenberg JuddH, Tham T (2020) The prevalence of olfactory and gustatory dysfunction in COVID-19 patients: a systematic review and meta-analysis. Otolaryngol Head Neck Surg (US) 163(1):3–11. https://doi.org/10.1177/0194599820926473
Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K (2020) SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J Hum Genet 65(12):1075–1082. https://doi.org/10.1038/s10038-020-0808-9
Trunfio M, Venuti F, Alladio F, Longo BM, Burdino E, Cerutti F, Ghisetti V et al (2021) Diagnostic SARS-CoV-2 cycle threshold value predicts disease severity, survival, and six-month sequelae in COVID-19 symptomatic patients. Viruses 13(2):2–14. https://doi.org/10.3390/v13020281.
Tsui BCH, Deng A, Pan S (2020) COVID-19: epidemiological factors during aerosol-generating medical procedures. Anesth Analg. https://doi.org/10.1213/ANE.0000000000005063
Tutuncu EE, Ozgur D, Karamese M (2021) Saliva samples for detection of SARS-CoV-2 in mildly symptomatic and asymptomatic patients. J Med Virol 93(5):2932–2937. https://doi.org/10.1002/jmv.26821
van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ et al (2020) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83(April):104351. https://doi.org/10.1016/j.meegid.2020.104351
World Health Organization (2021) WHO COVID-19 case definition. https://www.who.int/publications/i/item/WHO-2019-nCoV-Surveillance_Case_Definition-2020.2. Accessed 3 June 2021
Yang J, Zheng Ya, Gou Xi, Ke Pu, Chen Z, Guo Q, Ji R, Wang H, Wang Y, Zhou Y (2020) Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: a systematic review and meta-analysis. Int J Infect Dis 94:91–95. https://doi.org/10.1016/j.ijid.2020.03.017
Zoabi Y, Deri-Rozov S, Shomron N (2021) Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Digit Med 4(1):1–5. https://doi.org/10.1038/s41746-020-00372-6