BMJ Public Health

Clinical characteristics of COVID-19 in children and adolescents: insights from an Italian paediatric cohort using a machine-learning approach

Stefania Fiandrino,[1,2] Daniele Donà,[3,4] Carlo Giaquinto,[3,4] Piero Poletti,[5] Michael Davis Tira,[4] Costanza Di Chiara ,[3,4] Daniela Paolotti[2]

To cite: Fiandrino S, Donà D, Giaquinto C, *et al*. Clinical characteristics of COVID-19 in children and adolescents: insights from an Italian paediatric cohort using a machine-learning approach. *BMJ Public Health* 2025;**3**:e001888. doi:10.1136/bmjph-2024-001888

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjph-2024-001888).

DD, CG, PP, MDT, CDC and DP contributed equally.

Received 13 August 2024
Accepted 30 April 2025

[1]University of Rome La Sapienza, Rome, Italy
[2]ISI Foundation, Torino, Italy
[3]Department of Women's and Children's Health, Università degli Studi di Padova, Padua, Italy
[4]Penta Foundation, Padua, Italy
[5]Fondazione Bruno Kessler, Trento, Italy

Correspondence to
Dr Costanza Di Chiara; costanza.dichiara@phd.unipd.it

ABSTRACT

Introduction The epidemiology and clinical characteristics of COVID-19 evolved due to new SARS-CoV-2 variants of concern (VOCs). The Omicron VOC's higher transmissibility increased paediatric COVID-19 cases and hospital admissions. Most research during the Omicron period has focused on hospitalised cases, leaving a gap in understanding the disease's evolution in community settings. This study targets children with mild to moderate COVID-19 during pre-Omicron and Omicron periods. It aims to identify patterns in COVID-19 morbidity by clustering individuals based on symptom similarities and duration of symptoms and develop a machine-learning tool to classify new cases into risk groups.

Methods We propose a data-driven approach to explore changes in COVID-19 characteristics by analysing data from 581 children and adolescents collected within a paediatric cohort at the University Hospital of Padua. First, we apply an unsupervised machine-learning algorithm to cluster individuals into groups. Second, we classify new patient risk groups using a random forest classifier model based on sociodemographic information, pre-existing medical conditions, vaccination status and the VOC as predictive variables. Third, we explore the key features influencing the classification through the SHapley Additive exPlanations.

Results The unsupervised clustering identified three severity risk profile groups. Cluster 0 (mildest) had an average of 1.2 symptoms (95% CI 0.0 to 5.0) and mean symptom duration of 1.26 days (95%CI 0.0 to 9.0), cluster 1 had 2.27 symptoms (95% CI 1.0 to 6.0) lasting 3.47 days (95% CI 1.0 to 12.0), while cluster 2 (strongest symptom expression) exhibited 3.41 symptoms (95% CI 2.0 to 7.0) over 5.52 days (95% CI 0.0 to 16.0). Feature importance analysis showed that age was the most important predictor, followed by the variant of infection, influenza vaccination and the presence of comorbidities. The analysis revealed that younger children, unvaccinated individuals, those infected with Omicron and those with comorbidities were at higher risk of experiencing a greater number and longer duration of symptoms.

Conclusions Our classification model has the potential to provide clinicians with insights into the children's risk profile of COVID-19 using readily available data. This approach can support public health by clarifying disease burden and improving patient care strategies. Furthermore, it underscores the importance of integrating risk classification models to monitor and manage infectious diseases.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ COVID-19 has shown evolving epidemiology and clinical characteristics due to emerging SARS-CoV-2 variants, with research primarily focusing on hospitalised cases or the early Omicron period, leaving gaps in understanding the disease's progression in community settings.

WHAT THIS STUDY ADDS

⇒ This study introduces a machine-learning model that identifies patterns in COVID-19 morbidity and predicts infection type and disease progression based on patient profiles.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY:

⇒ The approach can improve patient management by guiding clinical decisions and support public health efforts by providing a clearer understanding of disease burden, potentially influencing both clinical practice and public health policy.

INTRODUCTION

The epidemiology and clinical characteristics of COVID-19 evolved during the pandemic, largely due to the emergence of new SARS-CoV-2 variants of concern (VOCs) with different virulence and transmissibility. These changes in VOCs contributed to shifts in COVID-19 clinical manifestations and morbidity. The emergence of the B.1.1.529 (Omicron) VOC has been marked by a predominance of upper respiratory tract symptoms, such as rhinitis, cough and sore throat, resulting in a lower incidence of severe outcomes among adults. However, the

higher transmissibility of the Omicron VOC, combined with school reopening,[1] has led to a rise in paediatric COVID-19 cases,[2–4] significantly increasing hospital admissions among children[5] and, consequently, severe outcomes in absolute terms.

Although the WHO has declared an end to COVID-19 as a public health emergency,[6] SARS-CoV-2 continues to persist and mutate. Coupled with a significant decline in global vaccination uptake and coverage, the risk remains of new VOCs emerging, potentially causing new surges in cases and deaths.

Given that the clinical characteristics of COVID-19 vary with different viral strains, understanding and early recognition of SARS-CoV-2 infection in the paediatric population is crucial to reducing the global burden of the pandemic.[7 8] With the decline in testing, providing evidence on the clinical patterns of paediatric COVID-19 is essential for facilitating early recognition and prompt management of cases.

To date, most research describing the changing symptomatology of COVID-19 during the Omicron period has concentrated on hospitalised cases, focusing on more severe cases and limiting our understanding of the disease's evolution in community settings, which represent the majority of cases.[4 9 10]

This research focuses on the youngest population infected with mild to moderate COVID-19, covering pre-Omicron and Omicron infections from April 2020 to December 2022 in the Veneto region of Italy. The analysed data consist of records of children aged 0–20 years seeking care from family paediatricians (FPs). The study aims to achieve two primary objectives: (1) uncovering patterns in COVID-19 morbidity by clustering individuals according to the number and duration of symptoms experienced; (2) developing a machine-learning tool to classify new cases based on demographic data, treatments and coexisting medical conditions and vaccination status, using the classes of infection identified in the previous step.

This study builds on prior research by Di Chiara *et al*,[11] which investigated the epidemiological and clinical features of SARS-CoV-2 variants using descriptive statistics. Our research aims to reinforce these findings by employing an unsupervised machine-learning approach to analyse clinical manifestations in children. This approach helps clinicians to understand the classes of SARS-CoV-2 infections, thus the children's risk profiles, and the possible burden of disease, facilitating better decision-making and personalised treatment.[9 12 13]

METHOD

Dataset description

In this study, we rely on data collected within a prospective cohort of 715 participants focusing on children and adolescents aged 0–21 years old attending the COVID-19 Family Cluster Follow-up Clinic (CovFC) from April 2020 to December 2022.[11] The CovFC was instituted at the Department of Women's and Children's Health, University Hospital of Padua, situated in the Veneto region, Italy. Families, including children, older siblings and parents, who had recovered from COVID-19 were referred to the CovFC by their FPs, and to be eligible for the enrolment they had to meet two criteria: (1) have children under the age of 15 and (2) have one or more family members with a confirmed history of laboratory-confirmed COVID-19 infection. During enrolment, paediatricians and/or infectious diseases specialists conducted clinical assessments, including the collection of demographic information, medical history, SARS-CoV-2 virological test results from nasopharyngeal swabs and vaccination status.[14] Clinical assessments and data collection were conducted for all individuals, including both parents and children, regardless of their laboratory-confirmed COVID-19 history. Following this, individuals with confirmed COVID-19 cases underwent a 6-monthly clinical and serological follow-up for at least 1 year after the initial infection, while subjects who were asymptomatic and had no analytical evidence of SARS-CoV-2 infection were considered non-COVID-19 cases. Vaccination data were recorded as they became available for each age group. Two blinded paediatricians determined the baseline infection date for each individual in the study, as outlined in Di Chiara *et al*.[11]

In the current study, we implement additional specific exclusion criteria. Specifically, individuals classified as non-COVID-19 cases, and those older than twenty years, were excluded from the analysis. Within the sample of 715 participants, 119 individuals were classified as non-COVID-19 cases, and 15 individuals were aged more than twenty years. Following the exclusion criteria, we discard those cases from the dataset, resulting in a final dataset including 581 children and adolescents.

Variables definition

We analyse data on existing medical conditions, vaccination status and reported symptoms in the paediatric population, gathered through clinical assessments conducted at the enrolment. In terms of existing medical conditions, we first check the prevalence of each in the study sample, removing the ones without any representation in the dataset, and then we consider those among the list of comorbidities associated with severe paediatric COVID-19: chronic pulmonary conditions (eg, bronchopulmonary dysplasia and uncontrolled asthma); cardiovascular conditions (eg, congenital heart disease); immunocompromising conditions (eg, malignancy, primary immunodeficiency and immunosuppression); neurological conditions (eg, epilepsy and select chromosomal/genetic conditions); prematurity; feeding tube dependence and other pre-existing technology dependence requirements; diabetes mellitus; obesity.[15–18]

To include vaccination status information, we deal with missing values reported in the dataset for the vaccination against COVID-19 due to the availability of the vaccines in the study period. For this reason, we consider

the approval releases per age group: individuals older than 12 years old are considered vaccine-eligible from 31 May 2021,[19] while individuals aged 5–11 years old are vaccine-eligible from 1 December 2021[20] and children younger than 4 years old were vaccine-ineligible when the enrolment was open. The individuals infected before the approval date of the vaccine were classified as non-vaccinated, and individuals aged 0–4 years old are all considered not vaccinated. Within the symptom set, non-verbal symptoms for the younger age group, such as headache and small-taste alterations, have been excluded from the analysis. However, symptoms recognisable by parents, including myalgia and abdominal pain, have been retained. The final set includes fever, rhinitis, cough, dyspnoea, myalgia, arthralgia, sore throat, conjunctivitis, asthenia, abdominal pain, nausea, lack of appetite, skin rash, confusion, ear pain and other symptoms.

Starting from the available data, we extract additional information including the total number of symptoms reported during infection, the total number of comorbidities, the length of each symptom, the median duration of symptoms, the variant of infection and the hexavalent vaccination (diphtheria-tetanus-acellular pertussis, Polio, Hib and hepatitis B). Specifically, we define an infection category for each individual, considering three types of infection: asymptomatic (duration of symptoms=0 days), short (duration of symptoms≤5 days) and long infection (duration of symptoms >5 days). These categories were defined in consultation with paediatricians who participated in the enrolment process. To identify specific variants of infections, we consider that from a clinical and immunovirological point of view, the Parental and Delta variants exhibited striking similarities. With the emergence of the Omicron variant, marked by substantial mutations in the spike-receptor binding domain (S-RBD), a notable shift in the clinical, immunological and epidemiological aspects of COVID-19 occurred. For these reasons, we classified cases into two groups based on the reported baseline date of infection onset: pre-Omicron and Omicron, defining any SARS-CoV-2 infection occurring before 15 November 2021, as pre-Omicron, and infections occurring after that day as Omicron. Finally, to include information on vaccination history, we combine available information on individual vaccinations and the hexavalent vaccination variable to determine whether an individual has received multiple vaccines intended to protect against several diseases (diphtheria, tetanus, pertussis (DTP), inactivated/oral poliovirus vaccine (IPV/OPV), hepatitis B virus (HBV), *haemophilus influenzae* type b (Hib)).

Study population

This study examines 581 children and adolescents who tested positive for SARS-CoV-2 (COVID-19 cases, symptomatic), aged 0–20 years old. The dataset includes sociodemographic and health-related information. The 66.5% of the study population (386 individuals) infected by SARS-CoV-2 were older than 5 years old, while the gender was balanced. Most of the subjects do not show previous underlying disease: only 25% of the entire study population exhibit at least one medical condition among the ones associated with severe paediatric COVID-19. During the pre-Omicron phase, almost all individuals had not done the COVID-19 vaccination yet, probably due to the vaccine eligibility. At the same time, during the Omicron variant, the number of vaccinated children and adolescents increased (46 individuals out of 135 individuals infected during Omicron). As regards symptoms, only 7% of the infected people during the Omicron variant report no symptoms, while more than 70% present at least two symptoms. On the contrary, during the pre-Omicron period, nearly 35% of the individuals reported no symptoms, and less than 40% of the infected presented two or more symptoms. We provide a summary of clinical and sociodemographic characteristics (see table 1), including counts, percentages, medians and IQRs, as applicable. The stratification is based on the distinct phases considered, pre-Omicron and Omicron. To better characterise the cohort, we check the prevalence of comorbidities and symptoms (see online supplemental figure S1 and S2). Nearly 75% do not exhibit any comorbidities, followed by individuals with other comorbidities, asthma, prematurity and congenital heart disease. Lots of comorbidities included during the reporting phase do not show any representation in the dataset. For this reason, we remove them from the analysis, together with comorbidities not associated with severe COVID-19 in the paediatric age, that emerge to also have a low prevalence in the dataset (chronic hepatitis, rheumatic disease, nephropathy, haematological disease). We finally consider nine comorbidities: asthma, prematurity, congenital heart disease, neurological disease, diabetes, chronic respiratory disease, obesity and others. The most common symptoms are fever and rhinitis, followed by headache, asthenia and cough. We do not consider both headache and smell and taste alterations for the analysis to avoid biases, as they are non-verbal symptoms for the youngest population.

Unsupervised clustering

To identify underlying patterns and structures within the dataset, we employ a clustering approach, an unsupervised machine-learning technique that groups data elements based on inherent statistical similarities, without requiring predefined class labels.[21] The objective is to derive insights into the risk stratification of SARS-CoV-2 infections in children by leveraging the presence or absence of symptoms and the duration of infection. The clustering process generates class labels that characterise individuals according to the similarity of their reported symptoms, spanning seventeen distinct symptom types encoded as binary variables, as well as their classification within one of three infection duration categories.

We employ the K-modes algorithm,[22] an extension of the well-established K-Means algorithm. K-means is well known for its efficiency in clustering large data sets. However, its limitation to numeric data restricts its

Table 1 Overview of sociodemographic and clinical characteristics in the study population, stratified by Omicron-infected and pre-Omicron group based on the SARS-CoV-2 variant

	Pre-Omicron (N=428)	Omicron (N=135)	Total (N=581)
Age–median (IQR 25–75)	8.0 (4.0–11.0)	8.0 (5.0–10.0)	8.0 (4.0–11.0)
Age 0–2 years—No (%)	74 (17.29)	18 (13.33)	96 (16.52)
Age 3–5 years—No (%)	71 (16.59)	24 (17.78)	99 (17.04)
Age 6–10 years—No (%)	143 (33.41)	60 (44.44)	207 (35.63)
Age 11–13 years—No (%)	98 (22.90)	26 (19.26)	129 (22.20)
Age 14–20 years—No (%)	42 (9.81)	7 (5.19)	50 (8.61)
Gender male—No (%)	201 (46.96)	61 (45.19)	266 (45.95)
Gender female—No (%)	227 (53.04)	74 (54.81)	314 (54.05)
At least one comorbidity—No (%)	99 (23.13)	45 (33.33)	146 (25.13)
COVID-19 vaccination, done—No (%)	6 (1.40)	46 (34.07)	53 (9.12)
COVID-19 vaccination, not done - No (%)	422 (98.60)	89 (65.93)	528 (90.88)
No of symptoms—median (IRQ 25–75)	1.0 (0.0–2.0)	2.0 (1.0–3.0)	1.0 (0.0–2.0)
No of symptoms 0—No (%)	137 (32.01)	7 (5.19)	163 (23.45)
No of symptoms 1—No (%)	115 (26.87)	28 (20.74)	171 (30.66)
No of symptoms 2—No (%)	83 (19.39)	41 (30.37)	122 (23.00)
No of symptoms more than 2—No (%)	93 (21.73)	59 (43.70)	125 (21.52)
Median duration infection—median (IRQ 25–75)	0.0 (0.0–2.0)	1.0 (0.0–3.0)	0.0 (0.0–2.0)
Infection category: asymptomatic—No (%)	246 (57.48)	65 (48.15)	387 (66.61)
Infection category: short infection—No (%)	127 (29.67)	57 (42.22)	166 (21.69)
Infection category: long infection—No (%)	55 (12.85)	13 (9.63)	68 (11.71)

applicability in fields such as data mining, where extensive categorical datasets are commonly encountered. Addressing the challenge of clustering large categorical datasets in data mining, Huang and Ng introduced the K-modes algorithm. This algorithm is a modified version of the K-means in which a simple matching dissimilarity measure tailored for categorical variables instead of the Euclidean distance is employed. Unlike K-means, the K-modes algorithm uses modes instead of means for clusters and a frequency-based approach to update modes during the clustering process.[23 24] The clustering procedure requires specifying the number of groups to divide individuals a priori. To determine the optimal number of clusters, we use the Elbow method, which is based on the principle that increasing the number of clusters initially leads to a rapid reduction in total within-cluster variance, followed by a gradual levelling off. The optimal number of clusters is identified at the point where the decrease in variance slows significantly, forming an 'elbow' in the graph.[25] Specifically, we calculate the total within-cluster sum of squares for cluster values ranging from 1 to 6. The point where the graph sharply bends, beyond which further increases in clusters yield minimal reduction in variance, indicates the optimal K value.

The analysis was performed using the kmodes library in Python (V.0.12.2, https://github.com/nicodv/kmodes) within a custom analysis pipeline developed by our team, available at https://github.com/sfiandrino/clustering_symptoms_VERDI.

Classification model

To predict the risk class to which a newly diagnosed individual should be assigned, we employ a machine-learning classification model. These approaches have become increasingly popular in recent years for their ability to support better-individualised treatments. The risk groups are identified and defined by the output of the clustering. The predictive variables have been defined in consultation with paediatricians and include sociodemographics, vaccination status, comorbidities and variant of infection information. In the following, we report the extensive list: age, gender, asthma, prematurity, obesity, diabetes, chronic respiratory disease, congenital heart disease, neurological disease, presence of at least one comorbidity, COVID-19 vaccination, influenza vaccination and hexavalent vaccination, pre-Omicron/Omicron period of infection.

For classification, we use the random forest classifier, a supervised machine-learning algorithm that combines multiple decision trees. Each tree in the forest casts a vote, and the class with the majority of votes is selected as the most probable label for a given input.[26] This ensemble method is widely used for its speed, robustness to noise, and success in identifying non-linear patterns in

data. Also, it can effectively handle both numerical and categorical data, and it is resistant to overfitting.[27]

However, the analysed dataset is imbalanced, with the majority of samples belonging to clusters 0 and 1, while cluster 2 is underrepresented. This imbalance can cause the classifier to favour the majority classes, leading to poor performance on the minority class and reducing the model's overall generalisability. To address this issue, we apply the Synthetic Minority Over-sampling Technique for Nominal and Continuous data (SMOTENC), a modified version of SMOTE[28] which generates synthetic samples for both categorical and numerical features. For numerical features, SMOTENC uses Euclidean distance to identify nearest neighbours, while categorical features are assigned based on the most frequent category among those neighbours. This approach improves class balance in the training data, enhancing model learning without overfitting through data duplication. To prevent data leakage and bias, we apply SMOTENC only to the training set, ensuring that the test set remains representative of real-world conditions.

To further improve model performance and reduce bias, we employ a fivefold cross-validation strategy. Additionally, we use a grid search optimisation approach to systematically tune hyperparameters, selecting the optimal combination from a predefined set of alternatives. The list of hyperparameters and their corresponding values used in the grid search are provided in the online supplemental table S1.

For the classification task, we used the scikit-learn library in Python (V.1.5.1, https://scikit-learn.org/), implemented within our analysis pipeline.

Explainability

Machine-learning approaches are often perceived as black boxes, offering recommendations without revealing the underlying processes. Therefore, interpreting the results, understanding the hidden patterns and comprehending the reasoning behind the model's conclusion play a key role, especially when model outputs are used to support decision-making. To interpret the prediction model's output, we use the SHAP (SHapley Additive exPlanations) framework.[29] SHAP provides a unified measure of feature importance, aiming to understand each instance's prediction by quantifying the contribution of each feature. Originating from cooperative game theory, the Shapley value addresses the issue of determining each player's importance to the overall cooperation. Since features contribute to the model's output as players with varying magnitudes and signs, Shapley values consider both the magnitude and direction of their contributions[30] and enable the visualisation of the range and distribution of impacts on the model's output.[31]

SHAP analyses have been successfully applied to understand factors associated with COVID-19 outcomes, including mortality and severity risks,[32–34] admissions to intensive care units (ICUs) or emergency departments,[35 36] and mental health impacts on COVID-19 patients.[37] Specifically, Smith and Alvarez used SHAP values to identify mortality factors in COVID-19 patients, highlighting age, days in the hospital, lymphocytes and neutrophils as key variables influencing patient outcomes.[32] Rajwa *et al* used similar approaches to explore predictive patient characteristics for in-hospital mortality in COVID-19 cases.[34] Laatifi *et al* showed how the plasma level of cytokines in the blood influences the severity of SARS-CoV-2 infection.[33] Cavallaro *et al* and Duckworth *et al* applied SHAP to identify factors linked to ICU admissions and death in hospitalised patients,[35] and to emergency department attendance.[36] Finally, Ikram *et al* leveraged SHAP to identify predictors of poor mental health outcomes among adult Asian Indians affected by COVID-19.[37] However, the magnitude of SHAP values does not necessarily correspond to any causal relationships between a specific feature and the output.[38] Instead, it provides a powerful tool for understanding which features are most predictive of clinical conditions by highlighting correlations between features and outcomes, which can serve as a basis for generating testable hypotheses regarding potential causal mechanisms.

RESULTS

Characterisation of clusters attributes

To determine the optimal number of clusters, we apply the Elbow method, which identifies three distinct groups. The corresponding visualisation is provided in online supplemental figure S3. Following this, we conduct statistical analyses to uncover underlying patterns, structures, similarities within the groups and key differences among them. The results show that the three clusters correspond to distinct levels of total symptom count and median infection duration. Notably, this information was not incorporated during the clustering process, but emerged as a key finding, emphasising the value of the machine-learning approach in identifying meaningful patterns within the data. The clusters can be characterised as follows: Cluster 0 represents individuals exhibiting few or no symptoms, suggesting a higher likelihood of asymptomatic infection; cluster 1 and cluster 2 include individuals with a higher number of symptoms; cluster 2 includes COVID-19 cases with a longer likely duration of symptoms than people belonging to cluster 1. Table 2 shows the descriptive statistics per cluster, together with the analysis of variance one-way analysis to find which variables had a statistically different mean value between (at least two of those) the clusters. More statistics related to symptoms activity in different clusters are reported in online supplemental table S2. Cluster 1 and cluster 2 differ in the similarity of reported symptoms, in particular for fever, rhinitis and cough, and for the duration of the first two symptoms. Cluster 0 differs from the other two clusters because it captures the asymptomatic COVID-19 cases. Figure 1 shows the histogram of the percentage of individuals per number of symptoms. We find distinct patterns: within cluster 0, there is a notable

Table 2 Overview of cluster characteristics

Clinical characteristics	Cluster 0 (N=378)	Cluster 1 (N=113)	Cluster 2 (N=90)	ANOVA (p value)
Asymptomatic infection (No (%))	312 (82.54)	0 (0.00)	7 (7.78)	0.000
Short infection (No (%))	43 (11.38)	97 (85.84)	54 (60.00)	0.000
Long infection (No (%))	23 (6.08)	16 (14.16)	29 (32.22)	0.000
Median duration symptoms (mean (SD))	1.26 (7.37)	3.47 (7.36)	5.52 (13.3)	0.000
Number of symptoms (mean (SD))	1.2 (1.34)	2.27 (1.50)	3.41 (1.20)	0.000

Individuals are grouped based on the similarity of reported symptoms and infection category. For the infection categories, the table presents the number of individuals and the corresponding percentage within the group, while the average and SD are provided for symptom duration and the number of symptoms.
ANOVA, analysis of variance.

prevalence of individuals reporting no symptoms or a limited number of symptoms, while cluster 1 and cluster 2 show no representation among individuals reporting no symptoms; conversely, the behaviour reverses for the occurrence of a high number of symptoms, where cluster 1 and cluster 2 are prominent, while cluster 0 displays an opposing trend.

Classification process

We use a random forest classifier to predict the risk group of a new individual based on sociodemographic and clinical data. The results of the model yield a receiver operating characteristic (ROC) score of 0.78 (95% CI 0.74 to 0.81), indicating that our model, on average, effectively classifies 78% of cases into the defined classes. Figure 2 shows the confusion matrix, a visual representation of the actual vs predicted values, that measures the performance of the classification model. We report the raw confusion matrix and the row-wise normalised version to better understand the percentage of correct classifications and errors across classes. The diagonal represents correctly classified instances, and off-diagonal elements represent misclassifications. Specifically, people with few or no symptoms (cluster 0) were correctly classified for 55% of cases (95% CI 49% to 58%) and misclassified as belonging to cluster 1 for 26% of cases (95% CI 21% to 34%). Individuals belonging to cluster 1 were correctly classified for 64% of cases (95% CI 50% to 76%) and misclassified as belonging to cluster 0 for 21% of cases (95% CI 12% to 29%). Finally, COVID-19 cases in cluster 2 were correctly classified for 66% of cases (95% CI 62% to 71%). Notably, when the model makes errors, it tends to misclassify individuals into the adjacent severity group rather than the more distant one. This pattern indicates that the model retains some discriminatory power, as it rarely assigns

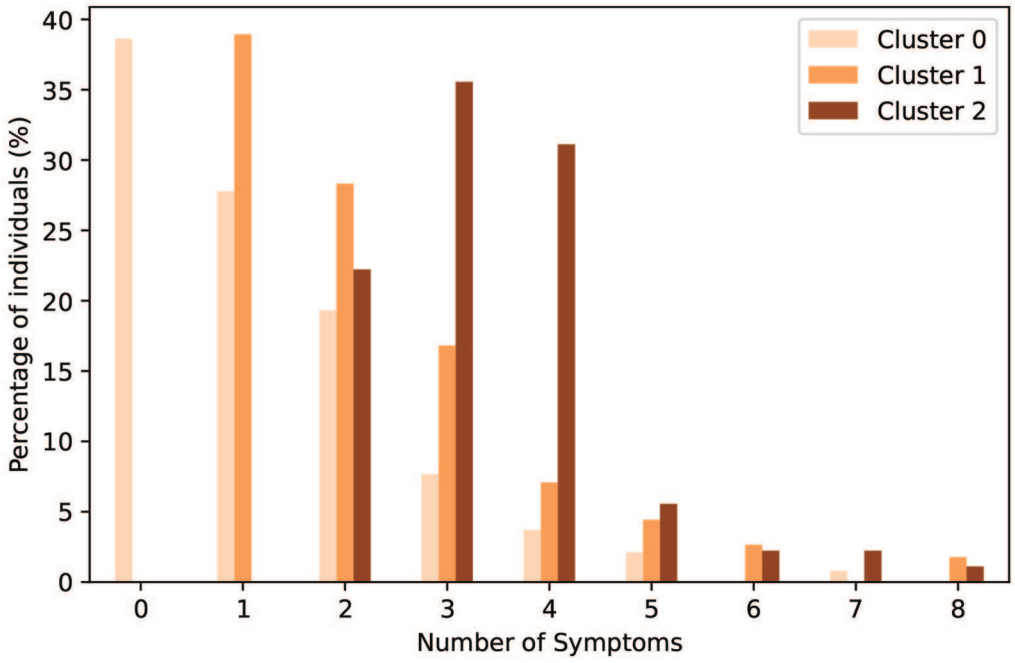

Figure 1 The figure presents the histogram of the prevalence of the number of symptoms reported by individuals within the three obtained clusters. Different patterns can be observed: individuals reporting no symptoms are assigned to cluster 0, while people reporting a higher number of symptoms belong to cluster 1 and cluster 2.

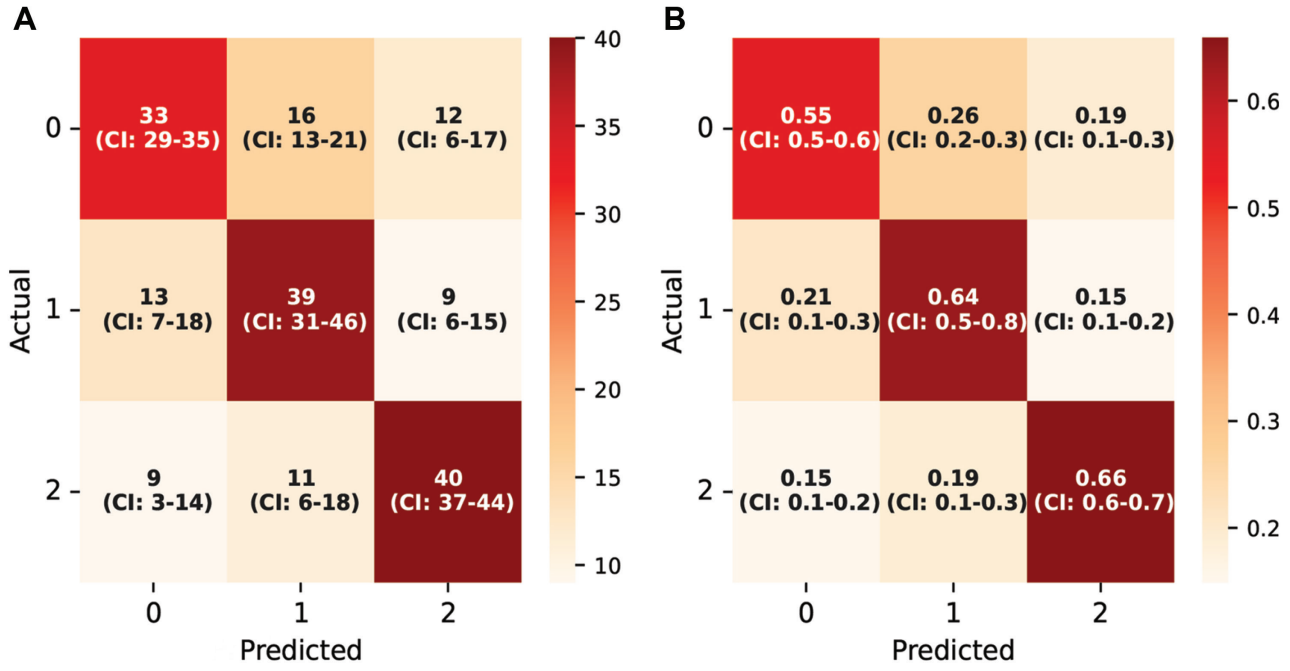

Figure 2 The figure presents the results of the multiclass classification model in distinguishing three severity risk groups (cluster 0 to cluster 2, ordered by increasing severity). (A) displays the raw confusion matrix, while (B) shows the row-wise normalised version, highlighting classification performance across classes. The x-axis represents the predicted labels, and the y-axis denotes the actual labels.

individuals from cluster 0 directly to cluster 2 or vice versa. Instead, misclassifications are more likely to occur between neighbouring clusters, reflecting the severity levels. Also, considering our three-class framework, a purely random classifier would achieve an accuracy of approximately 33%. In contrast, even the lower bounds of the CIs in our results do not approach this threshold, indicating that our model provides significant predictive value beyond random chance.

Explanation of predicted variables

To understand the most predictive variables for the random forest classifier, we use SHAP values.

Figure 3A reports the SHAP summary plot, where features are first sorted by their global impact, and dots represent the shape values, coloured by the value of that feature, from low (blue) to high (red). In other words, positive SHAP values indicate the feature increases the probability of assignment to the highest-risk group. Age appears to be the most important factor, and the colouring shows a smooth decrease in the model's output as age increases. Notably, we have similar results for gender, influenza vaccination and the presence of at least one comorbidity, meaning that a lower-risk profile characterises females, people with influenza vaccination and those without comorbidities. On the contrary, a higher risk profile characterises people infected during the Omicron variant, as shown by the opposite dot colour distribution. Figure 3 reports the mean absolute SHAP value of the features for the three classes, providing a general overview of the most influential features for the model (on the top) and their impact on the classification of each class.

The top five predominant factors identified as crucial for the classification task are age, the VOC, the presence of influenza vaccination, gender and the presence of at least one comorbidity. In addition to influenza vaccination, COVID-19 vaccination also emerged as a predictive factor, although with lower importance. This is likely due to the relatively low coverage among children in the analysed period, which may have limited its impact on the model's predictions.

DISCUSSION

The study employs a data-driven approach to characterise COVID-19 manifestations in children and adolescents during pre-Omicron and Omicron periods. Through an unsupervised machine-learning approach, we identify three clusters of individuals based on symptom type and duration, revealing distinct clinical patterns that emerged from the clustering process. Additionally, we develop a classification model using sociodemographic variables, vaccination status and the VOC to predict an individual's risk profile.

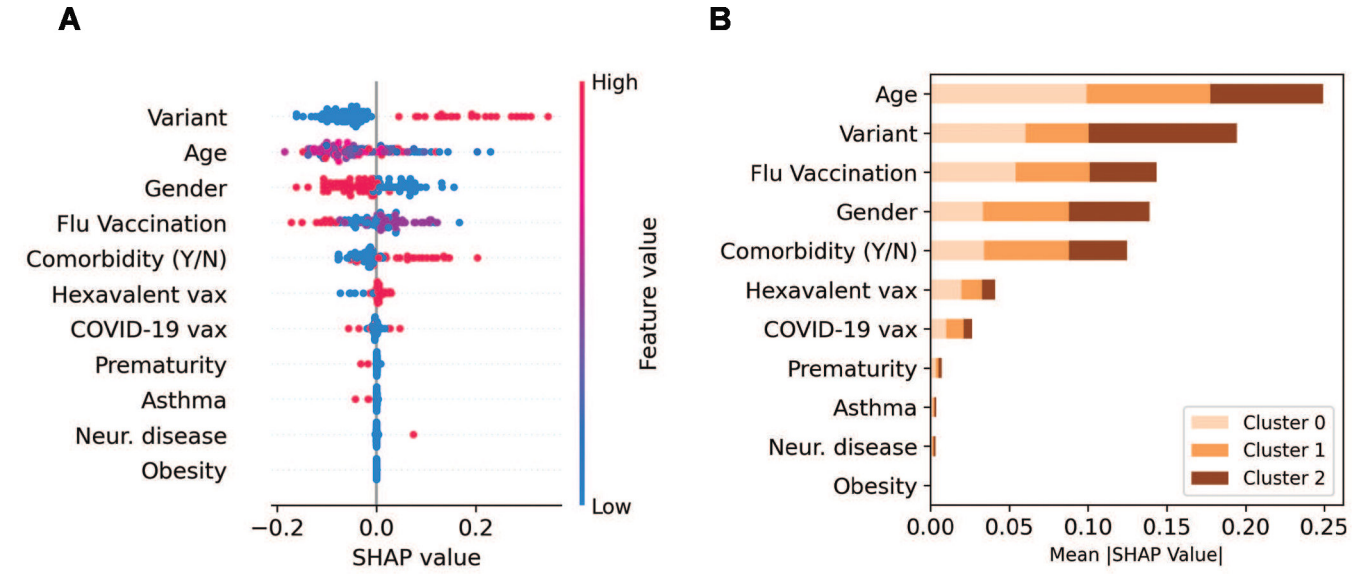

Figure 3 The figure presents the results of the explainability analysis using SHAP values: on the left (A) instance-individual SHAP values showing the impact on model output (higher risk group output), with importance ranking of the top variables and positive SHAP values indicating the feature increases the probability of assignment to higher risk group; on the right (B) global features importance based on the mean absolute magnitude of the SHAP values per class. SHAP, SHapley Additive exPlanations.

Our analysis confirms the findings reported by Di Chiara *et al*,[11] where statistical and clinical descriptive approaches were employed, showing a different pattern of clinical manifestation of COVID-19 in the paediatric population according to age, comorbidities, vaccination and infective VOC.

The unsupervised clustering approach identifies three risk profile clusters that reflect a spectrum of burden and infection duration: cluster 0, characterising individuals with fewer symptoms and most asymptomatic infections; cluster 1, characterising medium levels of number of symptoms and duration of symptoms; cluster 2, including the most symptomatic cases. This challenges the traditional binary classification of symptomatic versus asymptomatic cases and underscores the continuum of COVID-19 manifestations in children. Notably, symptom duration—though not an input variable—aligned with the identified clusters, highlighting the ability of machine learning to uncover clinically relevant patterns.

Clinically, the distinction between clusters 1 and 2 suggests that specific symptoms, such as fever, rhinitis and cough, may be key indicators of disease progression. These findings could help refine monitoring strategies, particularly in identifying children at risk of prolonged illness who may benefit from closer follow-up.

The increased frequency of upper respiratory tract symptoms in Omicron-infected children observed in our study is consistent with previous reports documenting a shift from lower to upper respiratory involvement.[2 4 11 39] This shift may explain the higher incidence of croup during the Omicron wave and has implications for clinical management, particularly regarding airway support in younger children.[40]

The classification model achieves moderate predictive performance (ROC-AUC=0.78), with age, VOC, vaccination status, gender and comorbidities emerging as the most influential predictive features. The impact of age and comorbidities aligns with previous studies, which have identified these factors as key determinants of COVID-19 severity.[41–43] The lower symptom burden observed in vaccinated children aligns with adult studies showing reduced systemic symptoms postimmunisation, therefore, confirming the role of vaccination in mitigating disease severity also in children.[44–46] This result highlights the importance of ongoing vaccination efforts, especially in light of the rapid expansion of SARS-CoV-2 variants and sublineages.[44–46] Interestingly, influenza vaccination emerged as a predicted variable for milder COVID-19, suggesting a potential cross-protection of influenza vaccination against COVID-19, as well as a potential enhanced effectiveness of dual influenza and COVID-19 seasonal vaccination in the paediatric population.[47 48] Interestingly, the model demonstrates a tendency to misclassify individuals into adjacent severity clusters rather than distant ones, reflecting the continuum of disease severity rather than discrete categories. This suggests that while our model captures relevant clinical patterns, further refinement—potentially incorporating additional clinical parameters—could enhance its precision.

Implications and applications

Despite COVID-19 becoming endemic among other seasonal respiratory viruses, the risk of new VOCs emerging with different virulence and transmissibility profiles, potentially leading to more severe cases, still persists. This work aids public health preparedness efforts

and clinical decision-making. Furthermore, recent years have shown significant changes in the epidemiology and clinical presentation of seasonal respiratory viruses, with more severe cases of influenza and RSV among older children.[49 50] In this context, a model that predicts infection type and progression based on a patient's profile can guide clinical decisions, improving patient management and outcomes. As self-diagnosis becomes more common, it is crucial to recognise the limitations of self-diagnosis in terms of specificity and sensitivity, which can lead to misdiagnosis. Supplementing testing with clinical insights is essential to accurately identifying severity and risk profiles. The rise of self-testing also brings the risk of overtreatment, especially the overuse of antibiotics, which is a global health threat due to antibiotic resistance. A precise risk profile model can support clinicians in distinguishing between infections, helping to reduce unnecessary antibiotic prescriptions at the community level. Moreover, this model could be particularly beneficial in low-income and middle-income countries where resources are limited. The ability to classify risk and predict disease progression using minimal resources can aid healthcare providers in these regions, improving patient outcomes.

Strengths and limitations

Using data from a prospective cohort ensured more accurate and consistent data collection, limiting reporting bias. However, the present work comes with some limitations. The framework needs further testing on a substantially larger dataset, including the integration of socioeconomic information and the most severe cases such as hospitalised patients with the need for medical care (eg, oxygen, ventilatory support). Similar to influenza, given the numerous variables that influence the risk of COVID-19 and the severity of the resulting illness, confounding is a significant issue in studies examining risk factors for COVID-19. Key potential confounders in these studies include socioeconomic variables such as household crowding, education level and income.[13] Nonetheless, despite the limited size of our study population, the focus on mild and moderate cases, and the missing information on more detailed socioeconomic aspects, we have identified differences in clinical manifestations among cases, highlighting distinct infection classes.

CONCLUSIONS

This data-driven approach provided different risk profile classes of COVID-19 in children using readily available information such as clinical history, VOC, vaccination status and sociodemographic factors. This helps predict the risk profile group for a new patient. Overall, our findings highlight the importance of integrating risk-classification models to improve the management of infectious diseases, not only for COVID-19 but also for other respiratory infections. Further research is needed to profile classes of COVID-19 severity in children. This approach can support public health efforts by providing a clearer understanding of disease burden and facilitating better resource allocation and patient care strategies.

X Costanza Di Chiara @C_DiChiara_MD

Acknowledgements The corresponding author would like to thank Dr Bertilla Ranzato for her support in patient enrolment. The authors thank all the family paediatricians collaborating with the project. The authors thank all families who attended the COVID-19 family clusters follow-up Clinic of the University Hospital of Padova.

Contributors SF conceptualised and designed the study, performed the analysis and wrote the manuscript. DD performed the patients' enrolment investigations, data curation, interpretation and visualisation and contributed to the writing. CG conceptualised and supervised the study and contributed to the writing. PP performed the validation and methodology and contributed to the writing. MDT performed the validation and methodology and contributed to the writing. CDC conceptualised and designed the study, performed the patients' enrolment and investigations, data curation and interpretation, supervised the study and contributed to the writing. DP conceptualised and designed the study and methodology, supervised the study and contributed to the writing. CDC and DP contributed equally as co-last authors. All authors had full access to all the data in the study, approved the final manuscript as submitted and accepted responsibility for submitting it for publication. CDC and DP are the guarantors of this study and take full responsibility for the integrity of the work, from inception to publication. Both authors ensure that all aspects of the work, including study design, data collection, analysis and interpretation, have been conducted and reported accurately and ethically.

Funding This work is part of the VERDI project (101045989), which is funded by the European Union.

Disclaimer Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants. The study protocol was approved by the Ethics Committee of the University Hospital of Padova, Italy (Prot. No 0070714 of 24 November 2020; last amendment Prot. No 0024018 of 4 May 2022). Parents/legally authorised representatives were informed of the research proposal and provided written consent to participate in the study and use the collected patient data. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Requests should be made to the corresponding author.

ORCID iD
Costanza Di Chiara http://orcid.org/0000-0002-3586-0612

REFERENCES

1 Bassi F, Doria M. Diffusion of COVID-19 among children and adolescents during the second and third waves of the pandemic in Italy. *Eur J Pediatr* 2022;181:1619–32.
2 Taytard J, Prevost B, Schnuriger A, *et al*. SARS-CoV-2 B.1.1.529 (Omicron) Variant Causes an Unprecedented Surge in Children Hospitalizations and Distinct Clinical Presentation Compared to the SARS-CoV-2 B.1.617.2 (Delta) Variant. *Front Pediatr* 2022;10:932170.
3 Han MS, Kim KM, Oh KJ, *et al*. Distinct Clinical and Laboratory Features of COVID-19 in Children During the Pre-Delta, Delta and Omicron Wave. *Pediatr Infect Dis J* 2023;42:423–8.
4 Westerhof I, de Hoog M, Ieven M, *et al*. The impact of variant and vaccination on SARS-CoV-2 symptomatology; three prospective household cohorts. *Int J Infect Dis* 2023;128:140–7.
5 Curatola A, Ferretti S, Graglia B, *et al*. COVID-19 increased in Italian children in the autumn and winter 2021-2022 period when Omicron was the dominant variant. *Acta Paediatr* 2023;112:290–5.
6 Wise J. Covid-19: WHO declares end of global health emergency. *BMJ* 2023;381:p1041.
7 Nyberg T, Ferguson NM, Nash SG, *et al*. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. *Lancet* 2022;399:1303–12.
8 Han L, Shen P, Yan J, *et al*. Exploring the Clinical Characteristics of COVID-19 Clusters Identified Using Factor Analysis of Mixed Data-Based Cluster Analysis. *Front Med (Lausanne)* 2021;8:644724.
9 Quintero AM, Eisner M, Sayegh R, *et al*. Differences in SARS-CoV-2 Clinical Manifestations and Disease Severity in Children and Adolescents by Infecting Variant. *Emerg Infect Dis* 2022;28:2270–80.
10 Aiello TF, Puerta-Alcalde P, Chumbita M, *et al*. Infection with the Omicron variant of SARS-CoV-2 is associated with less severe disease in hospitalized patients with COVID-19. *J Infect* 2022;85:e152–4.
11 Di Chiara C, Boracchini R, Sturniolo G, *et al*. Clinical features of COVID-19 in Italian outpatient children and adolescents during Parental, Delta, and Omicron waves: a prospective, observational, cohort study. *Front Pediatr* 2023;11:1193857.
12 Cui X, Zhao Z, Zhang T, *et al*. A systematic review and meta-analysis of children with coronavirus disease 2019 (COVID-19). *J Med Virol* 2021;93:1057–69.
13 Gordon A, Reingold A. The Burden of Influenza: a Complex Problem. *Curr Epidemiol Rep* 2018;5:1–9.
14 Di Chiara C, Cantarutti A, Costenaro P, *et al*. Long-term Immune Response to SARS-CoV-2 Infection Among Children and Adults After Mild Infection. *JAMA Netw Open* 2022;5:e2221616.
15 Woodruff RC, Campbell AP, Taylor CA, *et al*. Risk Factors for Severe COVID-19 in Children. *Pediatrics* 2022;149:e2021053418.
16 Graff K, Smith C, Silveira L, *et al*. Risk Factors for Severe COVID-19 in Children. *Pediatr Infect Dis J* 2021;40:e137–45.
17 Farrar DS, Drouin O, Moore Hepburn C, *et al*. Risk factors for severe COVID-19 in hospitalized children in Canada: A national prospective study from March 2020-May 2021. *Lancet Reg Health Am* 2022;15:100337.
18 Please provide reference 18.
19 Agenzia Italiana del Farmaco. AIFA approva l'utilizzo del vaccino comirnaty per la fascia di età 12-15 anni. n.d. Available: https://www.aifa.gov.it/-/aifa-approva-l-utilizzo-del-vaccino-comirnaty-per-la-fascia-di-et%C3%A0-12-15-anni
20 Agenzia Italiana del Farmaco. AIFA approva il vaccino comirnaty per la fascia di età 5-11 anni. n.d. Available: https://www.aifa.gov.it/-/aifa-approva-il-vaccino-comirnaty-per-la-fascia-di-et%C3%A0-5-11-anni
21 Talabis MR, *et al*. Analytics defined. In: *Information Security Analytics*. Boston: Syngress, 2015: 1–12.
22 Chaturvedi A, Green PE, Caroll JD. K-modes Clustering. *J of Classification* 2001;18:35–55.
23 Goyal M. Granth Sahib World University Fatehgarh Sahib, India. A review on K-mode clustering algorithm. *Int J Adv Res Comput Sci* 2017;725–9.
24 Huang Z, Ng MK. A Note on K-modes Clustering. *J Classif* 2003;20:257–61.
25 Bholowalia P, Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *Int J Comput Appl* 2014;105.
26 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
27 Robnik-Šikonja M. Improving random forests. In: *Machine Learning: ECML 2004*. Berlin, Heidelberg: Springer, 2004: 359–70.
28 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic Minority Over-sampling Technique. *jair* 2002;16:321–57.
29 Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv; 2017.
30 Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J Comput Aided Mol Des* 2020;34:1013–26.
31 Lundberg SM, Erion G, Lee SI. Consistent individualized feature attribution for tree ensembles. 2018;arXiv.
32 Smith M, Alvarez F. Identifying mortality factors from Machine Learning using Shapley values - a case of COVID19. *Expert Syst Appl* 2021;176:114832.
33 Laatifi M, Douzi S, Ezzine H, *et al*. Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME. *Sci Rep* 2023;13:5481.
34 Rajwa B, Naved MMA, Adibuzzaman M, *et al*. Identification of predictive patient characteristics for assessing the probability of COVID-19 in-hospital mortality. *PLOS Digit Health* 2024;3:e0000327.
35 Cavallaro M, Moiz H, Keeling MJ, *et al*. Contrasting factors associated with COVID-19-related ICU admission and death outcomes in hospitalised patients by means of Shapley values. *PLoS Comput Biol* 2021;17:e1009121.
36 Duckworth C, Chmiel FP, Burns DK, *et al*. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Sci Rep* 2021;11:23017.
37 Ikram M, Shaikh NF, Vishwanatha JK, *et al*. Leading Predictors of COVID-19-Related Poor Mental Health in Adult Asian Indians: An Application of Extreme Gradient Boosting and Shapley Additive Explanations. *Int J Environ Res Public Health* 2022;20:775.
38 Ma S, Tourani E. Predictive and causal implications of using shapley value for model interpretation. In: Proceedings of the 2020 KDD Workshop on Causal Discovery. PMLR; 2020:23–38. Available: https://proceedings.mlr.press/v127/ma20a.html
39 Tagarro A, Coya O-N, Pérez-Villena A, *et al*. Features of COVID-19 in Children During the Omicron Wave Compared With Previous Waves in Madrid, Spain. *Pediatr Infect Dis J* 2022;41:e249–51.
40 Shoji K, Akiyama T, Tsuzuki S, *et al*. Clinical characteristics of COVID-19 in hospitalized children during the Omicron variant predominant period. *J Infect Chemother* 2022;28:1531–5.
41 Barek MA, Aziz MA, Islam MS. Impact of age, sex, comorbidities and clinical symptoms on the severity of COVID-19 cases: A meta-analysis with 55 studies and 10014 cases. *Heliyon* 2020;6:e05684.
42 Chen A, Huang J-X, Liao Y, *et al*. Differences in Clinical and Imaging Presentation of Pediatric Patients with COVID-19 in Comparison with Adults. *Radiol Cardiothorac Imaging* 2020;2:e200117.
43 Ludvigsson JF. Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults. *Acta Paediatr* 2020;109:1088–95.
44 Trevisan C, Noale M, Prinelli F, *et al*. Age-Related Changes in Clinical Presentation of Covid-19: the EPICOVID19 Web-Based Survey. *Eur J Intern Med* 2021;86:41–7.
45 Xue FX, Shen KL. COVID-19 in children and the importance of COVID-19 vaccination. *World J Pediatr* 2021;17:462–6.
46 Zimmermann P, Pittet LF, Finn A, *et al*. Should children be vaccinated against COVID-19? *Arch Dis Child* 2022;107:e1:e1–8:.
47 Xie Z, Hamadi HY, Mainous AG, *et al*. Association of dual COVID-19 and seasonal influenza vaccination with COVID-19 infection and disease severity. *Vaccine (Auckl)* 2023;41:875–8.
48 Del Riccio M, Lorini C, Bonaccorsi G, *et al*. The Association between Influenza Vaccination and the Risk of SARS-CoV-2 Infection, Severe Illness, and Death: A Systematic Review of the Literature. *Int J Environ Res Public Health* 2020;17:7870.
49 European Centre for Disease Prevention and Control. Intensified circulation of respiratory syncytial virus (rsv) and associated hospital burden in the eu/eea. 2022 Available: https://www.ecdc.europa.eu/en/publications-data/intensified-circulation-respiratory-syncytial-virus-rsv-and-associated-hospital
50 Tokars JI, Olsen SJ, Reed C. Seasonal Incidence of Symptomatic Influenza in the United States. *Clin Infect Dis* 2018;66:1511–8.