nature communications

Article

https://doi.org/10.1038/s41467-024-55273-z

Racial/ethnic differences in post-acute sequelae of SARS-CoV-2 in children and adolescents in the United States

Received: 22 March 2024

Accepted: 5 December 2024

Published online: 21 January 2025

Check for updates

Dazheng Zhang[1,2,19], Bingyu Zhang[1,3,19], Qiong Wu[1,2,4], Ting Zhou[1,2], Jiayi Tong[1,2,5], Yiwen Lu[1,3], Jiajie Chen[1,2], Huiyuan Wang[1,2], Deena J. Chisolm[6], Ravi Jhaveri[7], Rachel C. Kenney[8,9], Russell L. Rothman[10], Suchitra Rao[11], David A. Williams[12], Mady Hornig[13], Linbo Wang[14], Jeffrey S. Morris[2], Christopher B. Forrest[15] & Yong Chen[1,2,3,16,17,18]

Racial/ethnic differences are associated with the symptoms and conditions of post-acute sequelae SARS-CoV-2 infection (PASC) in adults. These differences may exist among children and warrant further exploration. We conducted a retrospective cohort study with difference-in-differences analyzes to assess these differences in children and adolescents under the age of 21. The study utilized data from the RECOVER Initiative in the United States, which aims to learn about the long-term effects of COVID-19. The cohort included 225,723 patients with SARS-CoV-2 infection or COVID-19 diagnosis between March 2020 and October 2022. The study compared minority racial/ethnic groups to Non-Hispanic White (NHW) individuals, stratified by severity during the acute phase of COVID-19. Within the severe group, Asian American/Pacific Islanders (AAPI) had a higher prevalence of fever/chills and respiratory signs and symptoms, Hispanic patients showed greater hair loss prevalence in severe COVID-19 cases, while Non-Hispanic Black (NHB) patients had fewer skin symptoms in comparison to NHW patients. Within the non-severe group, AAPI patients had increased POTS/dysautonomia and respiratory symptoms, and NHB patients showed more cognitive symptoms than NHW patients. In conclusion, racial/ethnic differences related to COVID-19 exist among PASC symptoms and conditions in pediatrics, and these differences are associated with the severity of illness during acute COVID-19.

The post-acute sequelae SARS-CoV-2 infection (PASC) has emerged as a significant concern[1–3], particularly among young individuals with a previous diagnosis of COVID-19 [4–7]. Defined by the World Health Organization (WHO) as the persistence of at least one physical symptom for 12 weeks following initial testing without an alternative diagnosis and expanded by the National Institutes of Health (NIH) to include ongoing, relapsing, or new symptoms four or more weeks post-acute infection, PASC potentially affects a significant proportion of COVID-19 survivors[8–10]. In pediatric populations, recent studies show that PASC symptoms and conditions tend to be systematic and/or syndromic, with higher incidence conditions such as loss of taste/smell, myocarditis, and symptoms associated with cold-like illness occurring in patients after the acute phase of COVID-19[11]. The estimated prevalence of PASC symptoms and conditions in the pediatric population ranges from 23% to 45% among those previously infected by SARS-CoV-2[12–14], depending on study designs and PASC definitions.

A full list of affiliations appears at the end of the paper. e-mail: ychen123@upenn.edu

These findings highlight the urgent need for further research and comprehensive support to address the prevalence of PASC in pediatric populations.

Prior investigations into potential racial/ethnic differences in PASC among adults have unearthed important findings. For example, the Centers for Disease Control and Prevention (CDC) have shown variations in PASC's impact based on race/ethnicity[15]. A study by Khullar et al[16]. reported that Non-Hispanic Black (NHB) individuals exhibited a higher incidence of new PASC symptoms and conditions compared to Non-Hispanic White (NHW) patients, a difference more pronounced in hospitalized than in non-hospitalized patients. These findings suggest the existence of potential racial/ethnic differences in PASC among adults. Importantly, it is crucial to note that race and ethnicity are social constructs rather than biological ones[17,18]. Concurrently, research indicates children's likelihood of testing positive for COVID-19 correlates with their race/ethnicity[19–22]. NHB, Hispanic, and multi-racial children exhibited higher rates of COVID-19 positivity compared to their NHW counterparts, indicating differences in infection rates across different racial/ethnic groups[23]. However, limited research to date has addressed potential racial/ethnic differences in PASC among children and adolescents, making it a pressing area of study. Therefore, our study aimed to quantify such racial/ethnic differences by conducting an association study involving children and adolescents, to determine if the observed patterns are consistent with the findings from studies conducted among adults.

Examining health outcomes through the lens of racial/ethnic differences in the context of COVID-19 runs the risk of pre-existing racial/ethnic differences being either overshadowed or underestimated. To address this, we employed a difference-in-differences (DiD) approach to disentangle the shifts in racial/ethnic differences before and after COVID-19. Our investigation centered on a pediatric cohort from the RECOVER electronic health records (EHR) database across thirteen institutions. This study focused on the pediatric population with six months of follow-up, aims to investigate racial and ethnic differences in PASC symptoms and conditions attributable to SARS-CoV-2 infection. The study also included the Asian American/Pacific Islanders (AAPI) population, which was absent in the prior research.

Our main aim is to ascertain whether COVID-19 status correlates with any racial/ethnic differences across PASC symptoms and conditions. We use the term PASC symptoms and conditions to include a broad spectrum of health issues observed following COVID-19 infection[11,24], which are not exclusive to the post-COVID period. This includes systemic manifestations such as fatigue and malaise, respiratory symptoms, cardiovascular complications, and neurological disorders. Our definition, broader than the clinical diagnosis of PASC (ICD-10 code U09.9), enables us to investigate prevalence changes across different racial/ethnic groups before and after COVID-19 infection. Importantly, we quantified the racial/ethnic differences linked with COVID-19 by carefully accounting for pre-infection disparities in PASC symptoms and conditions through the application of DiD analyses.

Specifically, Fig. 1a demonstrates how to use the DiD model to find the racial/ethnic differences attributable to COVID-19. We measured the pre-infection racial/ethnic differences by comparing PASC symptoms and conditions between minority racial/ethnic groups and the NHW group during the pre-infection period. We then measured post-infection differences using the same metrics. The racial/ethnic differences attributable to COVID-19 were determined by calculating the difference between these post-infection and pre-infection differences. When COVID-19 may modify the effect of unmeasured confounding variables, we applied a negative control method (Fig. 1c) to adjust for the shifted impacts of these potential unmeasured confounders. By addressing these questions, our objective is to shed light on racial/ethnic differences in the relationship between COVID-19 and PASC symptoms and conditions.

Results

Study cohort and characteristics

The study involved 225,723 children and adolescents with COVID-19 (COVID-19-positive cohort) from March 2020 to October 2022 (cohort entry month). The index date for the cohort was the earliest date of documented SARS-CoV-2 infection, determined by a positive polymerase chain reaction (PCR), serology, antigen test, or diagnosis of COVID-19 or PASC (U09.9). The COVID-19-positive cohort comprised 109,022 NHW patients (48.3%), 45,823 NHB patients (20.3%), 60,012 Hispanic patients (26.6%), and 10,866 AAPI patients (4.8%) with SARS-CoV-2 infection from March 2020 to October 2022 (Fig. 2). Of these, 50.2% were female.

We stratified the COVID-19-positive patients at cohort entry into two groups based on the severity during the acute phase (seven days before to fourteen days after the index date) of the COVID-19 infection[25]: the severe group (moderate and severe cases) and the non-severe group (asymptomatic and mild cases). Table 1 and Table S4 present the baseline characteristics of the cohort, such as age, sex, and Pediatric Medical Complexity Algorithm[26] (PMCA), stratified by race/ethnicity and severity. The definitions for each variable are shown in Table S2 of Supplementary Section S1.

Statistical method overview

The PASC symptoms and conditions were defined using ICD-10-CM codes (shown in Table S2) and were measured during the pre-infection period (28 to 179 days before the index date) and the post-infection period (28 to 179 days after the index date). We performed pairwise propensity score matching for each minority racial/ethnic group (NHB, Hispanic, AAPI) to the NHW group based on pre-specified covariates. To adjust for pre-infection racial/ethnic differences, we applied a DiD model with negative control outcome (NCO) calibration (Fig. 1c) to the matched cohort.

Figure 1 illustrates the rationale and methodology of our difference-in-differences (DiD) with the NCO calibration approach. Figure 1b shows the rationale of the DiD model. The parallel trends assumption in the DiD model within our study states that the racial/ethnic differences do not change over time in the absence of COVID-19 infection. When the parallel trends assumption holds, unmeasured confounding variables do not affect the racial/ethnic difference in outcomes before and after COVID-19 infection. This is represented by the absence of an arrow between the unmeasured confounding variables and the differences in outcomes in Fig. 1b. However, when the parallel trends assumption is violated—specifically if COVID-19 modifies the effect of unmeasured confounding variables—the outcome estimates from the DiD may become biased without the NCO calibration. To address this limitation, we implemented negative control experiments for each minority racial/ethnic group compared to the NHW group, as illustrated in Fig. 1c. This approach allows us to calibrate our estimates from the unmeasured confounding bias. We provide details for explanations for Fig. 1b and Fig. 1c in the "Methods" Section.

Racial/ethnic differences in PASC symptoms and conditions

To investigate racial/ethnic differences attributable to COVID-19, we examined twenty-four PASC symptoms and conditions. We grouped the PASC symptoms and conditions into two categories: systemic and syndromic features (shown in Table S3). Syndromic features encompass symptoms, signs, and non-specific laboratory abnormalities, while systemic features are diagnosed health conditions. We used thirty-one NCOs in this study, defined as outcomes assumed to have no racial/ethnic differences attributable to COVID-19 (Table S1).

After achieving the balance of standardized mean difference (SMD) by propensity score matching (Supplementary Section S11) and conducting the DiD model with NCO calibration, Fig. 3 shows the racial/ethnic differences attributable to COVID-19 in PASC symptoms and conditions stratified by the severity of acute COVID-19 after the

https://doi.org/10.1038/s41467-024-55273-z

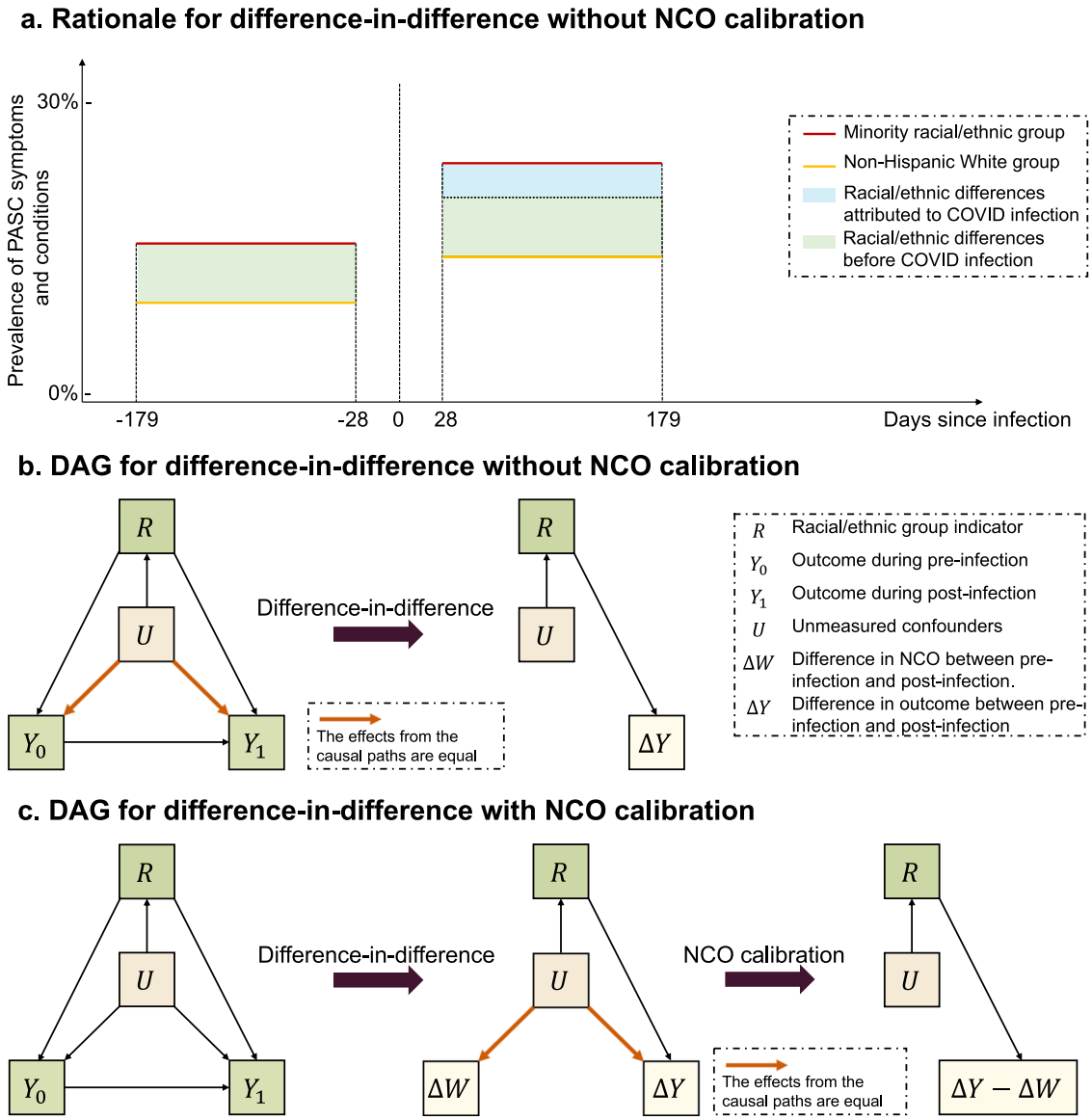

Fig. 1 | Illustration of difference-in-differences (DiD) analysis. a The rationale for DiD without the negative control outcome (NCO) calibration. It disentangles the racial/ethnic differences related to COVID-19 infection in PASC symptoms and conditions from the pre-infection observed racial/ethnic differences. **b** The directed acyclic graph (DAG) for the DiD without NCO calibration. Each node in the DAG represents each variable and the arrow symbol shows the causal effect. The left panel illustrates the parallel trends assumption for DiD. The right panel demonstrates how DiD without NCO calibration effectively blocks pathways from unmeasured confounders to PASC symptoms and conditions, provided the parallel trends assumption holds. **c** The DAG for the DiD with NCO calibration. The left panel illustrates a scenario where the parallel trends assumption is violated. The middle panel demonstrates that DiD without NCO calibration fails to block the pathway from unmeasured confounders to PASC symptoms and conditions. The right panel highlights that the DiD method with NCO calibration successfully eliminates the bias from unmeasured confounding.

NCO calibration. Overall, we found evidence of an increase in composite outcomes, i.e., at least one condition and any of the syndromic conditions, after SARS-CoV-2 infection among the AAPI group in both severe and non-severe COVID-19 group, but there was no evidence of increased racial differences among Hispanic and NHB groups.

For patients with severe COVID-19, AAPI patients showed a higher increase in at least one condition (risk ratio [RR] 1.24, 95% confidence interval [CI] 1.04 to 1.49, two-sided Wald test P = 0.019) and any of syndromic conditions (RR 1.22, 95% CI 1.01 to 1.47, P = 0.042) compared to NHW after SARS-CoV-2 infection. Both Hispanic patients (RR 0.99, 95% CI 0.91 to 1.08, P = 0.804) and NHB patients (RR 0.93, 95% CI 0.85 to 1.24, P = 0.147) showed a non-significant decrease in at least one condition as compared to NHW patients. For patients with non-severe COVID-19, AAPI patients showed an increase in at least one condition (RR 1.08, 95% CI 1.01 to 1.14, P = 0.015) and any of syndromic conditions (RR 1.08, 95% CI 1.01 to 1.08, P = 0.017) compared to NHW. Hispanic patients showed a non-significant increase in at least one condition (RR 1.01, 95% CI 0.98 to 1.04, P = 0.498), while NHB patients had a similar non-significant risk (RR 0.99, 95% CI 0.89 to 1.11, P = 0.915) in at least one condition.

However, there exist significant differences among all minority groups across several PASC symptoms and conditions after SARS-CoV-2 infection. For example, for patients with severe COVID-19, the increased prevalence of hair loss among Hispanic patients was greater (RR 2.62, 95% CI 1.06 to 6.49, two-sided Wald test P = 0.038) than the increased prevalence among NHW patients. The corresponding increase in the prevalence of fever and chills among AAPI was greater (RR 1.41, 95% CI 1.01 to 1.97, P = 0.045) than NHW. NHB patients had a smaller change in skin symptoms (RR 0.74, 95% CI 0.58 to 0.96, P = 0.021) than NHW patients. For patients with non-severe COVID-19, AAPI patients had a

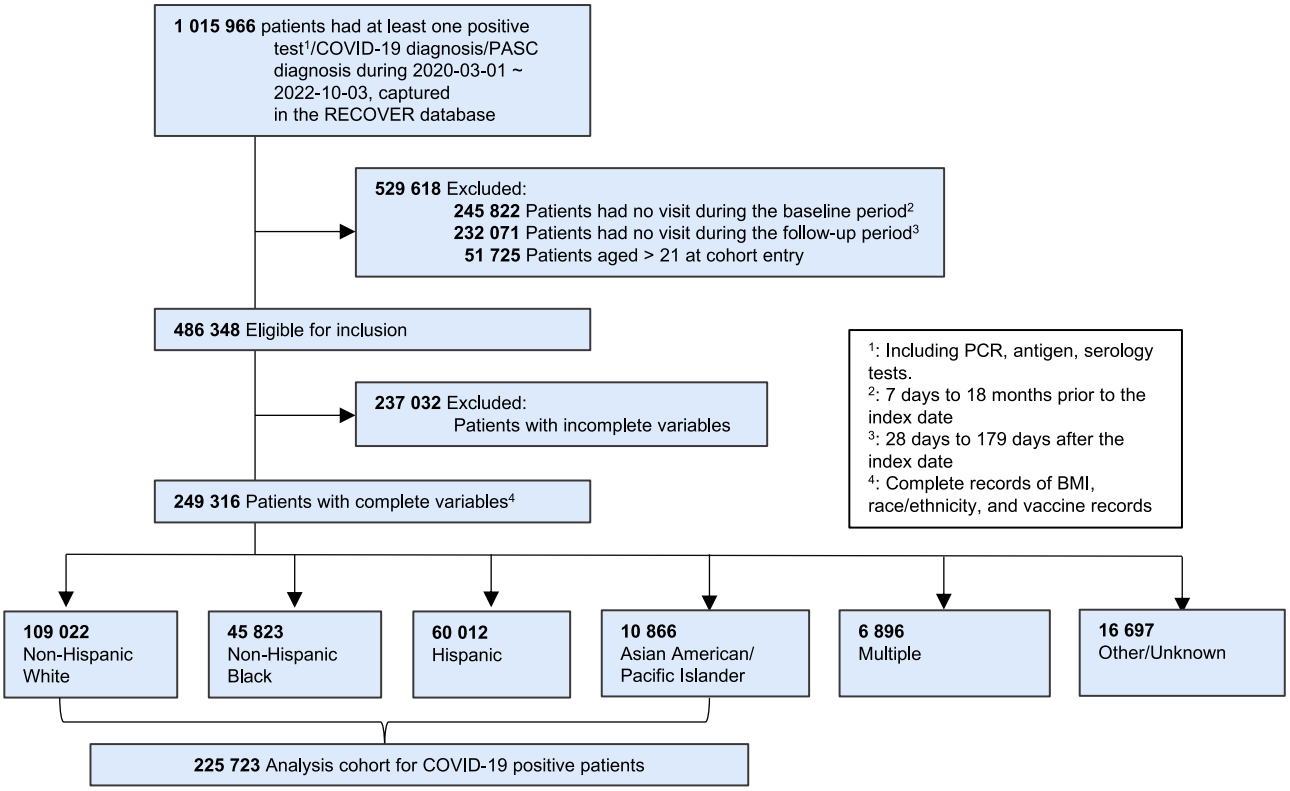

Fig. 2 | Flowchart for COVID-19 positive cohort. Flowchart illustrating the selection process for COVID-19-positive patients from March 2020 to October 2022. Abbreviation: BMI, body mass index; PASC, Post-Acute Sequelae of SARS-CoV-2 infection; PCR, Polymerase Chain Reaction.

greater change concerning POTS/dysautonomia (RR 1.57, 95% CI 1.02 to 2.40, P = 0.037) and respiratory signs and symptoms (RR 1.11, P = 0.036) with 95% CI 1.00 to 1.23 (significant; rounded to two decimals) compared to NHW patients. NHB patients had an increase in cognitive function (1.25, 95% CI 1.01 to 1.55, P = 0.037) than NHW patients.

Furthermore, we observed that these racial/ethnic differences varied depending upon the severity of the acute phase of COVID-19 as well as the specific PASC symptoms and conditions being analyzed. For example, among the severe group, the differential increase in abdominal pain was more pronounced for all three minority groups compared to those in the non-severe category.

Sensitivity analysis

Supplementary Section S2 showed the results of the NCO experiments and estimated systematic errors, such as the unmeasured confounder bias. Figure S14 in Supplementary Section S6 showed the racial/ethnic differences using standard Poisson regression models. Among COVID-19 patients within the severe group, NHB patients showed a greater incidence in at least one condition (RR 1.16, 95% CI 1.02 to 1.32, two-sided Wald test $P = 0.024$) and any syndromic conditions (RR 1.14, $P = 0.042$) with 95% CI 1.00 to 1.30 (significant; rounded to two decimals) as compared to NHW patients. Hispanic patients also showed a non-significant greater incidence in at least one condition (RR 1.12, 95% CI 0.99 to 1.27, $P = 0.075$) as compared to NHW patients.

In Fig. S14, among COVID-19 patients with severe illness during the acute infection, Hispanic individuals exhibited a greater incidence of respiratory signs and symptoms (RR 1.16, 95% CI 1.02 to 1.33, two-sided Wald test $P = 0.024$) and hair loss (RR 1.84, 95% CI 1.02 to 3.31, $P = 0.043$) as compared with the NHW patient group. NHB had a greater incidence of respiratory signs and symptoms (RR 1.19, 95% CI 1.03 to 1.36, $P = 0.015$) and heart disease (RR 1.48, 95% CI 1.06 to 2.07, $P = 0.022$), but a lower incidence of arrhythmias (RR 0.73, 95% CI 0.57 to 0.94, $P = 0.013$) and headache (RR 0.66, 95% CI 0.48 to 0.93, $P = 0.016$) compared with the NHW group. Among those with non-severe acute COVID-19, Hispanic patients displayed a higher incidence of myocarditis (RR 4.28, 95% CI 1.53 to 11.98, $P = 0.006$) and abnormal liver enzyme (RR 2.06, 95% CI 1.08 to 3.94, $P = 0.029$) compared with NHW patients. Meanwhile, AAPI patients demonstrated a greater incidence of hair loss (RR 3.32, 95% CI 1.43 to 7.72, $P = 0.005$) compared with the NHW patient group.

These findings revealed that our DiD approach identified fewer racial/ethnic differences compared to standard regression models. It is worth noting that the DiD approach adjusted for the baseline racial/ethnic differences before the SARS-CoV-2 infection, a step that a standard regression analysis failed to take into consideration. Consequently, some of the observed racial/ethnic differences with prior work might not be attributed to COVID-19. Nevertheless, given its adjustment for baseline racial/ethnic differences, the DiD approach holds greater robustness.

To determine whether the defined set of twenty-four PASC symptoms and conditions accurately reflect PASC, we conducted a crude incidence analysis comparing these symptoms and conditions in the COVID-19-positive cohort to that in the COVID-19-negative cohort (Supplementary Section S4). A total of 677,448 COVID-19-negative patients were included in the crude analysis. A random negative test date was chosen as the index date for COVID-19-negative patients and the selection criteria for the COVID-19-negative cohort are shown in Fig. S13. The results indicated that the incidence of all PASC symptoms and conditions was higher in all COVID-19-positive patients compared to all COVID-19-negative patients (Table S7).

To further validate our definition of PASC symptoms and conditions, as well as to investigate potential racial/ethnic differences in PASC symptoms attributable to COVID-19, we applied the DiD approach with NCO calibration to the COVID-19-negative cohort (Fig. S14). This analysis revealed no racial or ethnic differences in PASC symptoms and conditions attributable to COVID-19 within the

https://doi.org/10.1038/s41467-024-55273-z

Table 1 | Baseline characteristics of COVID-19 positive patients, by race/ethnicity and severity status

	Severe[a]				Non-severe[b]			
	NHW (N = 9,140)	AAPI (N = 590)	Hispanic (N = 3,786)	NHB (N = 3,346)	NHW (N = 99,882)	AAPI (N = 10,276)	Hispanic (N = 56,226)	NHB (N = 42,477)
Age categories								
< 1 years	1601 (17.5%)	122 (20.7%)	767 (20.3%)	733 (21.9%)	9655 (9.7%)	965 (9.4%)	6154 (10.9%)	4686 (11.0%)
1 to < 5 years	2811 (30.8%)	197 (33.4%)	1174 (31.0%)	910 (27.2%)	22018 (22.0%)	2802 (27.3%)	11892 (21.2%)	9147 (21.5%)
5 to < 12 years	1893 (20.7%)	133 (22.5%)	773 (20.4%)	568 (17.0%)	27547 (27.6%)	3191 (31.1%)	17205 (30.6%)	12991 (30.6%)
12 to < 16 years	1243 (13.6%)	60 (10.2%)	489 (12.9%)	482 (14.4%)	18276 (18.3%)	1433 (13.9%)	10430 (18.6%)	7721 (18.2%)
16 to < 21 years	1592 (17.4%)	78 (13.2%)	583 (15.4%)	653 (19.5%)	22386 (22.4%)	1885 (18.3%)	10545 (18.8%)	7932 (18.7%)
Sex								
Female	4208 (46.0%)	255 (43.2%)	1728 (45.6%)	1580 (47.2%)	50682 (50.7%)	4968 (48.3%)	28412 (50.5%)	21408 (50.4%)
Male	4932 (54.0%)	335 (56.8%)	2058 (54.4%)	1766 (52.8%)	49200 (49.3%)	5308 (51.7%)	27814 (49.5%)	21069 (49.6%)
Obesity								
No	6055 (66.2%)	422 (71.5%)	2230 (58.9%)	1926 (57.6%)	62750 (62.8%)	6141 (59.8%)	21525 (38.3%)	21711 (51.1%)
Yes	3085 (33.8%)	168 (28.5%)	1556 (41.1%)	1420 (42.4%)	37132 (37.2%)	4135 (40.2%)	34701 (61.7%)	20766 (48.9%)
PMCA								
No condition	4333 (47.4%)	311 (52.7%)	1938 (51.2%)	1552 (46.4%)	67424 (67.5%)	7594 (73.9%)	41294 (73.4%)	27868 (65.6%)
Non-complex chronic condition	1435 (15.7%)	79 (13.4%)	534 (14.1%)	524 (15.7%)	18614 (18.6%)	1564 (15.2%)	8820 (15.7%)	8692 (20.5%)
Complex chronic condition	3372 (36.9%)	200 (33.9%)	1314 (34.7%)	1270 (38.0%)	13844 (13.9%)	1118 (10.9%)	6112 (10.9%)	5917 (13.9%)
Negative tests prior entry								
0	6038 (66.1%)	407 (69.0%)	2671 (70.5%)	2301 (68.8%)	73195 (73.3%)	7908 (77.0%)	45852 (81.5%)	32592 (76.7%)
1	1445 (15.8%)	97 (16.4%)	521 (13.8%)	497 (14.9%)	15999 (16.0%)	1440 (14.0%)	6650 (11.8%)	6233 (14.7%)
2	641 (7.0%)	32 (5.4%)	225 (5.9%)	238 (7.1%)	5875 (5.9%)	519 (5.1%)	2169 (3.9%)	2146 (5.1%)
≥ 3	1016 (11.1%)	54 (9.2%)	369 (9.7%)	310 (9.3%)	4813 (4.8%)	409 (4.0%)	1555 (2.8%)	1506 (3.5%)
Vaccine dosage								
0	7860 (86.0%)	469 (79.5%)	3347 (88.4%)	3015 (90.1%)	82136 (82.2%)	7256 (70.6%)	46760 (83.2%)	37252 (87.7%)
1	240 (2.6%)	22 (3.7%)	80 (2.1%)	71 (2.1%)	2878 (2.9%)	474 (4.6%)	1667 (3.0%)	1159 (2.7%)
≥ 2	1040 (11.4%)	99 (16.8%)	359 (9.5%)	260 (7.8%)	14868 (14.9%)	2546 (24.8%)	7799 (13.9%)	4066 (9.6%)

NHW: Non-Hispanic White; NHB: Non-Hispanic Black; AAPI: Asian American or Pacific Islander; PMCA: Pediatric Medical Complexity Algorithm[26].
a: According to the definition of the severity of COVID-19[25] at cohort entry, the positive patients with moderate or severe COVID-19 were grouped into the severe group.
b: Similarly, the positive patients with asymptomatic or mild COVID-19 were grouped into the non-severe group.

COVID-19-negative cohort. These findings suggest that the observed racial/ethnic differences in PASC symptoms and conditions among COVID-19-positive individuals are unlikely to be due to inherent differences in symptom and condition incidence patterns but rather may reflect factors specifically associated with COVID-19 infection or its sequelae. This analysis validates that the racial and ethnic differences in PASC observed in the COVID-19-positive cohort are related to COVID-19 infection, rather than underlying population differences or unrelated health trends.

In the analysis including only those patients identified based on positive SARS-CoV-2 PCR or antigen testing (Supplementary Section S7), differences among severe patients were diminished among some PASC symptoms and conditions, while among the non-severe patients, the differences that we identified were consistent in both sets of analyses. To account for the potential bias from limited hospital capacity during the initial COVID-19 period, we performed a secondary analysis excluding COVID-19 patients from the first wave of the pandemic (March to May 2020). This exclusion did not significantly alter the results, as demonstrated in Supplementary Section S8. Supplementary Section S9 shows the results of subgroup analysis by age group. No significant differences were observed for patients aged < 5. For patients aged from 5 to 11, non-severe NHB patients showed an increase in at least one condition (RR 1.08, 95% CI 1.02 to 1.15, $P = 0.009$) and any of the syndromic conditions (RR 1.08, 95% CI 1.02 to 1.15, $P = 0.010$) compared to NHW group. For adolescents aged 12 to 20, differences among severe and non-severe patients had a similar pattern. Supplementary Section S10 shows the results of stratified analysis by virus variants. The Hispanic cohort showed less increase among severe and non-severe groups during the pre-Delta period; differences were consistent in both sets during the Delta period, and all minority groups showed increased risk in at least one condition and any syndromic conditions compared to NHW patients during the Omicron period. Supplementary Section S12 provided the results of adding the socioeconomic confounder Area Deprivation Index (ADI) into the analysis, and the differences detected were consistent in both sets of analyses.

Discussion

We examined racial/ethnic variations in long-term consequences of documented SARS-CoV-2 infection across thirteen health institutions in the RECOVER study for 225,723 children and adolescents. After accounting for pre-existing differences and potential confounding biases, we found evidence suggesting differences attributable to COVID-19 for AAPI patients compared to NHW patients. Notably, AAPI patients showed an increase in the risk of developing at least one PASC-related condition and any syndromic conditions, both in severe and non-severe COVID-19 cases. We observed specific symptom differences among minority racial/ethnic groups, despite no evidence indicating disparities in composite outcomes for NHB or Hispanic populations compared to NHW. For example, NHB patients showed a

https://doi.org/10.1038/s41467-024-55273-z

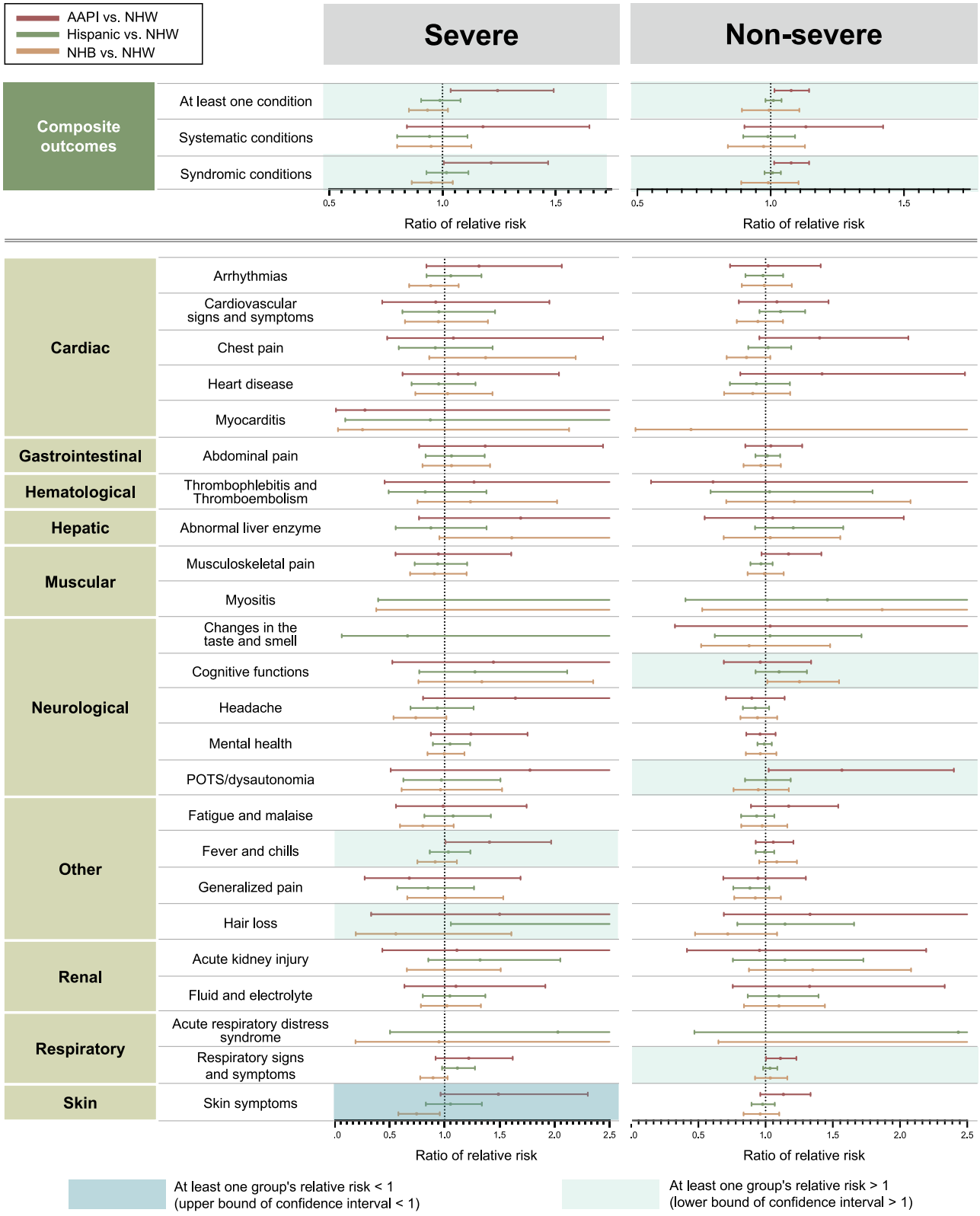

Fig. 3 | Racial/Ethnic differences attributable to COVID-19 in PASC symptoms and conditions. Racial/Ethnic differences in risk ratio (RR) that are attributable to COVID-19 for the prevalence of PASC symptoms and conditions among COVID-19-positive cohort (n = 225,723 participants), by race/ethnicity and severity status. Each minority group—Asian American/Pacific Islanders (AAPI), Non-Hispanic Black (NHB), and Hispanic—was compared to the Non-Hispanic White (NHW) group. The dots in the figure showed the estimated RRs for each minority group versus the NHW group from the difference-in-differences (DiD) analysis with NCO calibration stratified by severity status. The error bars showed the 95% confidence interval (CI) of the estimated RR. Source data are provided as a Source Data file.

smaller change in skin symptoms compared to NHW patients, consistent with previous findings in adults by Khullar et al[16].

The increased cognitive function issues in NHB patients, the greater prevalence of hair loss in Hispanic patients, the higher risk of POTS or dysautonomia, and respiratory symptoms in AAPI patients with severe COVID-19 all warrant particular attention in clinical settings. It is worth noting that AAPI is the smallest racial/ethnic group, and for this group, the results show evidence of differences compared to NHW. Therefore, this strengthens the findings of racial/ethnic differences across PASC symptoms and conditions for the other minority groups compared to NHW.

While the observed racial/ethnic differences in RRs are subtle, they remain significant when applied to large populations and may reflect broader systemic issues. For example, the observed differences in PASC can primarily be attributed to disparities in healthcare access and the availability of specialized pediatric care, which vary significantly across different regions of the country, according to a national report[27]. These variations are not due to biological reasons but are more closely linked to socioeconomic factors, healthcare infrastructure, and regional healthcare policies.

This study examines PASC racial/ethnic differences among pediatric populations, providing valuable insights into the clinical relevance of these disparities. Understanding these differences is crucial for developing targeted interventions to ensure equitable healthcare access and outcomes for all pediatric populations affected by PASC. The increased risk among minority racial/ethnic pediatric patients may necessitate targeted follow-up care and support for the minority racial/ethnic population. Healthcare providers should be aware of the potential for varied PASC presentations across racial/ethnic groups in pediatric populations. This knowledge can inform more targeted screening and follow-up protocols, potentially improving early detection and management of PASC symptoms in different racial/ethnic patient populations.

The implications of this study on a broader public health level are significant. By utilizing data from thirteen institutions, representing approximately 10% of the US children population and encompassing both urban and suburban health systems, this study offers a comprehensive national sample. Public health strategies are suggested to prioritize the reduction of healthcare access disparities and the enhancement of specialized pediatric care availability across all regions. This will help in better managing and treating PASC in children, ensuring that all affected patients receive the necessary care regardless of their geographic location or socioeconomic status. Furthermore, these findings underscore the importance of continuous surveillance, research, and tailored public health policies to address the evolving needs of this patient group.

The method utilized in our study has multiple strengths. First, we used propensity score matching methods instead of standard linear regression models in our adjustment of the confounders, which helped us reduce the non-linear effects of the confounders[28]. Second, we accounted for the pre-infection racial/ethnic differences in PASC symptoms and conditions. This approach enabled us to quantify racial/ethnic differences attributable to COVID-19 more accurately. The DiD model is particularly powerful in the context of studying racial/ethnic differences, as it helps to isolate the effect of COVID-19 from pre-existing health disparities. By comparing the change in health outcomes before and after COVID-19 across racial/ethnic groups, we can more confidently attribute observed differences to the impact of the virus rather than to the pre-infection racial/ethnic differences.

A key strength of our study is the implementation of NCOs, which are outcomes not expected to show racial/ethnic differences due to COVID-19. This approach enhances the reliability and interpretability of our findings. By analyzing these alongside our primary outcomes, we enable the detection of systematic bias from unmeasured confounding variables. This approach allows us to calibrate results from the DiD model. Importantly, NCOs help mitigate the impact of unmeasured confounding that may change before and after the infection. This is particularly crucial when studying racial/ethnic differences, where many dynamic social and environmental factors may not be captured in our dataset. Furthermore, we carefully addressed potential bias related to differences in follow-up times. To ensure comparability, we standardized the follow-up period for all participants, observing PASC symptoms and conditions over a consistent timeframe of five months (28 to 179 days post-cohort entry). This approach mitigates any bias that might arise from variations in follow-up duration across individuals.

This study has several limitations. First, socioeconomic differences may exacerbate racial/ethnic differences in PASC symptoms and conditions, thereby acting as a mediator effect in the causal pathway between race/ethnicity and clinical outcomes. Such influences have been suggested as risk factors for acute COVID-19 by Chisolm and colleagues in a RECOVER EHR study[23]. Future research on PASC outcomes is of interest to study such mediation effects.

Secondly, health-seeking behavior or healthcare access is an important consideration, known as ascertainment bias[29]. Limited access to care and associated medical records among certain minority racial/ethnic groups may contribute to potential bias in the observed racial/ethnic differences. Related issues were recently described by Nasir et al. for the ascertainment of PASC symptoms and conditions in adult populations through EHR[30]. Nevertheless, we tried to address the ascertainment bias by including healthcare utilization factors in the propensity score model, such as the number of inpatient, outpatient, and ED visits during the baseline period.

Thirdly, confounding poses a significant bias threat in observational studies. To address this, we extensively adjusted for potential confounders using a propensity-score-based matching method and DiD analyses. We employed NCOs to reduce the residual bias, such as unmeasured confounder bias. Additionally, EHR data completeness issues may lead to misclassification and loss-to-follow-up bias. Some attempts have been made to mitigate the impacts of these biases[31–34]. The analysis used a combined set of patients, but potential bias in racial/ethnic differences may vary between outpatients and inpatients. Addressing these issues can help improve the reliability of evidence generated from these investigations.

Fourthly, asymptomatic cases are less likely to be captured in EHR as they are less likely to be tested, impacting the findings for non-severe COVID-19. This study focused on diagnosed or treated cases of PASC, reflecting children who sought healthcare services. The direction of bias can be bidirectional, depending on whether asymptomatic cases are disproportionately missed in minority or NHW groups. This could attenuate the DiD analysis for the non-severe COVID-19 group due to potential misclassification bias, leading to conservative or inflated results. Future research should aim to capture a more comprehensive spectrum of COVID-19 severity.

Fifthly, our study design, requiring at least one healthcare visit within 28-179 days after the index date, balanced selection bias against loss to follow-up bias[35]. While this approach may overrepresent frequent healthcare users, especially during 2020-2021, we mitigated this by incorporating diverse testing data to reduce COVID-19 status misclassification. Furthermore, in the crude incidence analysis comparing COVID-19 positive and negative patients, the resulting selection bias likely included sicker patients in the COVID-19 negative group, potentially biasing our results towards the null. Nevertheless, we still observed a higher incidence of PASC symptoms and conditions for COVID-19-positive patients compared to COVID-19-negative patients. Future studies could explore alternative designs or data sources to enhance the external validity of these findings.

In summary, we quantified the racial/ethnic differences in PASC symptoms and conditions and the impact of SARS-CoV-2 infection on these differences. The impact of COVID-19 varied across racial/ethnic

groups, severity of acute COVID-19, and different PASC symptoms and conditions.

Methods

Ethics and inclusion

This study constitutes human subject research. Institute Review Board (IRB) approval was obtained under Biomedical Research Alliance of New York (BRANY) protocol #21-08-508. As part of the BRANY IRB process, the protocol has been reviewed in accordance with the institutional guidelines. The BRANY waived the need for patient-informed consent and HIPAA authorization.

Data source

This retrospective cohort study is part of the NIH Researching COVID-19 to Enhance Recovery (RECOVER) Initiative (https://recovercovid.org/), which aims to learn about the long-term effects of COVID-19. The data were contributed by thirteen sites. Participating institutions in this study included: Children's Hospital of Philadelphia, Cincinnati Children's Hospital Medical Center, Children's Hospital of Colorado, Ann & Robert H. Lurie Children's Hospital of Chicago, Nationwide Children's Hospital, Nemours Children's Health System (in Delaware and Florida), Duke University, University of Iowa Healthcare, University of Michigan, University of Missouri, Oregon Community Health Information Network (OCHIN), University of California, San Francisco, and Vanderbilt University Medical Center. All analyses were conducted using R version 4.1.2, with statistical significance set at a two-tailed p-value threshold of < 0.05.

Cohort construction

We conducted a retrospective study from March 1, 2020, to October 3, 2022, with at least 6 months of follow-up time. We included patients under the age of 21 who had at least one visit from 18 months before to 7 days before the index date (defined as the baseline period) and at least one encounter within 28 days and 179 days after the index date (defined as the follow-up period). For COVID-19-positive patients, we included the patients who had positive PCR, serology, or antigen tests or diagnoses of COVID-19, or diagnoses of PASC (U09.9), which we defined as documented SARS-CoV-2 infection. The index date for these patients was defined as the first time of SARS-CoV-2 infection. We used patient race/ethnicity data included in the EHR and categorized patients into six racial/ethnic groups: NHW, NHB, Hispanic, AAPI, Multiple, and Other/Unknown. Multiple comprising less than 3% of the population in any COVID-19 and severity category (Table S5 of Supplementary Section S3) were considered small samples and excluded from the analysis. The Other/Unknown categories were also excluded due to interpretability issues. The selection of participants for COVID-19-positive patients in real-world data is summarized in Fig. 2.

Defining outcomes

Our definition of PASC symptoms and conditions included 24 symptoms and conditions as shown in Rao et al[11]., including abdominal pain, abnormal liver enzyme, acute kidney injury, acute respiratory distress syndrome, arrhythmias, cardiovascular signs and symptoms, changes in taste and smell, chest pain, cognitive functions, fatigue and malaise, fever and chills, fluid and electrolyte, generalized pain, hair loss, headache, heart disease, mental health disorders, musculoskeletal pain, myocarditis, myositis, Postural Orthostatic Tachycardia Syndrome (POTS) or dysautonomia, respiratory signs and symptoms, skin symptoms, and thrombophlebitis and thromboembolism. We used validated diagnostic codes (ICD-10-CM) confirmed by board-certified pediatricians, with details of the code sets available in the Supplementary Materials Table S2. Systematic and syndromic conditions related to PASC were grouped by the 24 PASC symptoms and conditions, which are detailed in Table S4.

Patient characteristics

The primary exposure was race/ethnicity, categorized into NHW, NHB, Hispanic, and AAPI. Various patient characteristics were considered as confounders, such as age at the cohort entry date (< 5, 5–12, 12–21), sex (female, male), cohort entry month (from March 2020 to October 2022), site indicators, obesity (obese, non-obese), a chronic condition indicator as defined by the PMCA (no chronic condition, non-complex chronic condition, and complex chronic condition)[26], healthcare visits (inpatient, outpatient, and emergency department visits), medications (0, 1, 2, ≥ 3), negative tests (0, 1, 2, ≥ 3), vaccine doses (0, 1, ≥ 2), and immunization duration during the baseline period (no vaccine, < 4 months, ≥ 4 months). The severity of COVID-19 at the cohort entry date was stratified into the following categories: asymptomatic, mild (symptomatic), moderate (involving moderately severe COVID-19-related conditions like gastroenteritis, dehydration, and pneumonia), and severe (comprising unstable COVID-19-related conditions, ICU admission, or mechanical ventilation)[25]. We categorized patients exhibiting either asymptomatic or mild symptoms as belonging to the non-severe group, while all other patients were classified as part of the severe group.

Propensity score matching

To quantify the racial/ethnic differences in the PASC symptoms and conditions, we use RR as the comparative measure. The RR is known to be a collapsible measure, where collapsibility[36] refers to the measure of association conditional on some factors that remain consistent with the marginal measure collapsed over strata[37].

To eliminate the impact of potential measured confounders, we used a propensity score matching technique with the covariates detailed in the patient characteristics section. The propensity score is calculated by the logistic regression model fitted by regressing the racial/ethnic groups on the covariates. We performed this matching separately for minority racial/ethnic groups (NHB, Hispanic, and AAPI), each stratified by severity status, compared with the NHW group. After performing the matching, we assessed the SMD between each covariate value for different racial/ethnic groups, with a difference of 0.1 or less indicating an acceptable balance[38]. The detailed characteristics of the matched cohort are presented in Supplementary Section S5.

DiD models

Subsequently, we quantified the differential increase in the prevalence of PASC symptoms and conditions across different racial/ethnic groups by the DiD method on the matched cohort. Figure 1a provides a visual representation of the DiD method used to estimate racial/ethnic differences in the increased prevalence of PASC symptoms and conditions related to COVID-19. We introduce the notation to illustrate the DiD method. For each minority racial/ethnic group (AAPI, NHB, or Hispanic) matched with the NHW cohort, we use a binary indicator R to denote racial/ethnic groups ($R = 0$ for NHW, $R = 1$ for NHB, Hispanic, or AAPI). We define T as the time period, where $T = 1$ represents the post-infection period (28 to 179 days after COVID-19 infection) and $T = 0$ represents the pre-infection period (28 to 179 days before COVID-19 infection). Y is a binary indicator denoting the occurrence of PASC symptoms and conditions, where $Y = 1$ indicates the patient has been diagnosed with the specific PASC symptoms and conditions, and $Y = 0$ indicates the absence of such diagnosis. For each participant, we observed the outcome Y in both pre-infection ($T = 0$) and post-infection ($T = 1$) periods.

We fitted a Poisson regression model by regressing each PASC symptom or condition on racial/ethnic groups, time period, and the interaction between racial/ethnic groups and time period:

$$\log(E[Y|R,T]) = \beta_0 + \beta_1 R + \beta_2 T + \beta_3 R \times T \quad (1)$$

where log(·) is the log link function. From Eq. (1), the pre-infection racial/ethnic difference is found by the coefficient, β_1, shown in the green part in Fig. 1a. The post-infection racial/ethnic difference is denoted by the sum of the coefficients, $\beta_1+\beta_3$. Thus, the coefficient of the interaction term, β_3, represents the differential increase in racial/ethnic differences attributable to COVID-19, shown in the blue part in Fig.1a.

To illustrate the DiD model removes the bias from the unmeasured confounder when the parallel trends assumption holds, Fig. 1b shows the directed acyclic graph (DAG) for the causal relationship[39] when conducting the DiD model. We use Y_0 for the outcome during the pre-infection period, Y_1 for the outcome during the post-infection period, and U for the unmeasured confounding variables. For simplicity, we present the DAG for the matched cohort of each minority race/ethnicity-NHW pair after adjusting for measured confounding variables X using the matching method.

Arrows between variables denote causal relationships. For instance, the arrow from R to Y_1 represents the racial/ethnic difference during the post-infection period, while the arrow from R to Y_0 represents the difference during the pre-infection period. The DiD model aims to identify the racial/ethnic difference in the outcome change between the post-infection and the pre-infection period, ΔY (where $\Delta Y=Y_1-Y_0$), as shown by the arrow from R to ΔY.

The key advantage of the DiD model is its ability to adjust for unmeasured confounding variables U under the parallel trends assumption. This assumption states that the effect of U remains constant over time, meaning the effect of U on Y_0 and Y_1 is equal. In the DAG, this is represented by the path $U\rightarrow Y_0$ being equal to $U\rightarrow Y_1$.

Under this assumption, the DiD model, by taking the difference between Y_1 and Y_0, effectively blocks the path from U to ΔY, allowing us to isolate the effect of R on ΔY without interference from U. It is worth noting that a specified model with adjusting for the pre-infection outcomes Y_0 cannot adjust for the bias from the unmeasured confounder U. The model effectively blocks the path $R\rightarrow Y_0\rightarrow Y_1$. Additionally, it introduces new bias through the collider path $R\rightarrow Y_0\leftarrow U\rightarrow Y_1$. Collider bias occurs when two variables, such as R and U, influence a common effect, such as Y_0, and controlling for this effect inadvertently creates a spurious association between the variables, leading to biased results. This is why the DiD method is more accurate than the model adjusting for the outcome during the pre-infection period.

Negative control outcome calibration

While we used propensity score matching to account for the measured confounders and DiD analyses to address pre-infection racial/ethnic differences, the results can still be impacted by unmeasured confounder bias when the parallel trends assumption does not hold, that is, when COVID-19 modifies the effect of unmeasured confounding variables. To mitigate such bias, we collected thirty-one NCOs prespecified by pediatric physicians, which are not expected to show racial/ethnic differences due to COVID-19. We fitted the model from Eq. (1) with the outcome replaced by each NCO. The residual bias thus can be estimated from the average of β_3 s from the models for the NCOs. By using these NCOs, the study was able to calibrate the residual bias from unmeasured and systematic sources.

Figure 1c illustrates this by the expanded DAG. When the parallel trends assumption is violated, a backdoor path from U to ΔY can emerge, potentially biasing our estimates. To address this limitation, we incorporate NCOs into our analysis. Similarly, we denote each NCO during the post-infection period by W_1 and during the pre-infection period by W_0. Due to the influence of U, we observe a path from U to ΔW (where $\Delta W=W_1-W_0$). We then assume that the effect of U on ΔW is equivalent to its effect on ΔY. By leveraging this assumption, we can effectively block the path from U to ΔY-ΔW by subtracting ΔW from ΔY. This allows us to estimate the racial/ethnic difference in PASC symptoms and conditions with greater accuracy and robustness to unmeasured confounding. We conducted a two-sided Wald test to evaluate whether the racial/ethnic differences in PASC symptoms and conditions were equal to zero and reported the corresponding p-values. No adjustments were made for multiple comparisons.

Sensitivity analysis

We conducted a list of sensitivity analyses to examine the robustness of our findings. First, to evaluate how different statistical methods might influence the analytical results, we used an alternative approach, standard Poisson regression analyses, with RR as the comparative measure. Specifically, we considered the incidence of PASC symptoms and conditions as outcomes, while controlling for the same confounders that were used in the matching process within the DiD analysis. The p-values are reported from two-sided Wald tests. Second, we performed crude incidence analyses to compare PASC symptoms and conditions between COVID-19 positive and negative groups, assessing whether the prespecified outcomes accurately reflect PASC. We showed the p-values for the statistical significance from two-sided two-proportion Z-tests. In addition, to validate racial/ethnic differences attributed to COVID-19, we conducted DiD analyses with NCO calibration on COVID-19 negative groups. Third, we conducted analyses for COVID-19 patients identified only by positive SARS-CoV-2 PCR or antigen tests, because the recorded date of COVID-19 diagnosis may not accurately reflect the actual infection date. Fourth, we conducted analyses excluding patients whose index dates fell within the first wave of COVID-19 (March to May 2020) due to limited SARS-CoV-2 testing availability during this period. Additionally, our sensitivity analysis featured stratification by a set of age group strata (< 5, 5–11, 12–20), differing from the ones previously specified, and by estimated time frames corresponding to dominant virus variants (pre-Delta, Delta, Omicron). Finally, we conducted a sensitivity analysis including ADI as a measured confounder, which excluded patients who did not have this information.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Source data for figures are provided in this paper. The detailed individual-level patient data was compiled as part of the RECOVER program. Due to the high risk of reidentification based on the number of unique patterns in the data, patient privacy regulations prohibit us from releasing the data publicly. The data are maintained in a secure enclave, with access managed by the program coordinating center to remain compliant with regulatory and program requirements. Please direct requests to access the data, either for the reproduction of the work reported here or for other purposes, to the RECOVER EHR Pediatric Coordinating Center (recover@chop.edu). Source data are provided with this paper.

Code availability

The code used for the analysis in this study is available and can be accessed in the code ocean repository at https://doi.org/10.24433/CO.0460366.v1.

References

1. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol* **21**, 133–146 (2023).
2. Davis, H. E. et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine* **38**, 101019 (2021).
3. Thaweethai, T. et al. Development of a Definition of Postacute Sequelae of SARS-CoV-2 Infection. *JAMA* **329**, 1934–1946 (2023).

4. World Health Organization (WHO). A clinical case definition for post COVID-19 condition in children and adolescents by expert consensus. https://www.who.int/publications/i/item/WHO-2019-nCoV-Post-COVID-19-condition-CA-Clinicalcase-definition-2023-1 (2023).
5. Morello, R. et al. Risk factors for post-COVID-19 condition (Long Covid) in children: a prospective cohort study. *EClinicalMedicine* **59**, 101961 (2023).
6. Selvakumar, J. et al. Prevalence and characteristics associated With Post–COVID-19 condition among nonhospitalized adolescents and young adults. *JAMA Netw. Open* **6**, e235763 (2023).
7. Rodriguez-Morales, A. J. et al. The global challenges of the long COVID-19 in adults and children. *Travel Med. Infect. Dis.* **54**, 102606 (2023).
8. Lopez-Leon, S. et al. More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Sci. Rep.* **11**, 16144 (2021).
9. Groff, D. et al. Short-term and long-term rates of postacute sequelae of SARS-CoV-2 infection: a systematic review. *JAMA Netw. Open* **4**, e2128568 (2021).
10. Nasserie, T., Hittle, M. & Goodman, S. N. Assessment of the frequency and variety of persistent symptoms among patients with COVID-19: a systematic review. *JAMA Netw. Open* **4**, e2111417 (2021).
11. Rao, S. et al. Clinical features and burden of postacute sequelae of SARS-CoV-2 infection in children and adolescents. *JAMA Pediatr.* **176**, 1000–1009 (2022).
12. Twohig, H. et al. Long-term outcomes of COVID-19 infection in children and young people: a systematic review and meta-analysis. *medRxiv* 2023–2024 (2023).
13. Lopez-Leon, S. et al. Long-COVID in children and adolescents: A systematic review and meta-analyses. *Sci. Rep.* **12**, 9950 (2022).
14. O'Mahoney, L. L. et al. The prevalence and long-term health effects of long Covid among hospitalised and non-hospitalised populations: a systematic review and meta-analysis. *eClinical Med.* **55**, 101762 (2023).
15. National Center for Health Statistics, U.S. Census Bureau & Household Pulse Survey. Long COVID. https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm (2022).
16. Khullar, D. et al. Racial/ethnic disparities in post-acute sequelae of SARS-CoV-2 infection in New York: an EHR-based cohort study from the RECOVER program. *J. Gen. Intern Med.* **38**, 1127–1136 (2023).
17. Ford, M. E. & Kelly, P. A. Conceptualizing and categorizing race and ethnicity in health services research. *Health Serv. Res.* **40**, 1658–1675 (2005).
18. Ford, C. L. & Harawa, N. T. A new conceptualization of ethnicity for social epidemiologic and health equity research. *Soc. Sci. Med.* **71**, 251–258 (2010).
19. Saatci, D. et al. Association between race and COVID-19 outcomes among 2.6 million children in England. *JAMA Pediatr.* **175**, 928–938 (2021).
20. Swann, O. V. et al. Clinical characteristics of children and young people admitted to hospital with covid-19 in United Kingdom: prospective multicentre observational cohort study. *BMJ* **370**, m3249 (2020).
21. Bandi, S., Nevid, M. Z. & Mahdavinia, M. African American children are at higher risk of COVID-19 infection. *Pediatr. Allergy Immunol.* **31**, 861 (2020).
22. Lee, E. H. et al. Race/ethnicity among children with COVID-19–associated multisystem inflammatory syndrome. *JAMA Netw. Open* **3**, e2030280 (2020).
23. Chisolm, D. J. et al. Associations between area-level health-related social factor indices and risk of acute COVID-19: an EHR-based cohort study from the RECOVER program. *medRxiv* 2012–2022 (2022).
24. Razzaghi, H. et al. Vaccine effectiveness against long COVID in children. *Pediatrics* **153**, e2023064446 (2024).
25. Forrest, C. B. et al. Severity of acute COVID-19 in children<18 years old March 2020 to December 2021. *Pediatrics* **149**, e2021055765 (2022).
26. Simon, T. D., Haaland, W., Hawley, K., Lambka, K. & Mangione-Smith, R. Development and validation of the pediatric medical complexity algorithm (PMCA) version 3.0. *Acad. Pediatr.* **18**, 577–580 (2018).
27. Rivara, F. P., Gonzalez-del-Rey, J. & Forrest, C. B. The pediatric subspecialty physician workforce. *JAMA Pediatr.* **178**, 107–108 (2024).
28. Amoah, J. et al. Comparing propensity score methods versus traditional regression analysis for the evaluation of observational data: a case study evaluating the treatment of gram-negative bloodstream infections. *Clin. Infect. Dis.* **71**, e497 (2020).
29. Asch, D. A. et al. Variation in US hospital mortality rates for patients admitted With COVID-19 during the first 6 months of the pandemic. *JAMA Intern Med.* **181**, 471–478 (2021).
30. Nasir, M. et al. Using data science and a health equity lens to identify long-COVID sequelae among medically underserved populations. *J. Health Care Poor Underserved* **34**, 521–534 (2023).
31. Duan, R. et al. An empirical study for impacts of measurement errors on EHR based association studies. *AMIA Annu. Symp. Proc.* **2016**, 1764 (2016).
32. Chen, Y., Wang, J., Chubak, J. & Hubbard, R. A. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: empirical illustration using breast cancer recurrence. *Pharmacoepidemiol Drug Saf.* **28**, 264–268 (2019).
33. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc. Natl Acad. Sci.* **115**, 2571–2577 (2018).
34. Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A. & Madigan, D. Interpreting observational studies: why empirical calibration is needed to correct *p*-values. *Stat. Med.* **33**, 209–218 (2014).
35. Howe, C. J., Cole, S. R., Lau, B., Napravnik, S. & Eron, J. J. Jr Selection bias due to loss to follow up in cohort studies. *Epidemiology* **27**, 91–97 (2016).
36. Greenland, S., Pearl, J. & Robins, J. M. Confounding and collapsibility in causal inference. *Stat. Sci.* **14**, 29–46 (1999).
37. Whitcomb, B. W. & Naimi, A. I. Defining, quantifying, and interpreting "noncollapsibility" in epidemiologic studies of measures of "effect. *Am. J. Epidemiol.* **190**, 697 (2021).
38. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28**, 3083–3107 (2009).
39. Pearl, J. et al. *Causality*. (Cambridge University Press, 2009).

Acknowledgements

This study is part of the NIH Researching COVID-19 to Enhance Recovery (RECOVER) Initiative, which seeks to understand, treat, and prevent the post-acute sequelae of SARS-CoV-2 infection (PASC). For more information on RECOVER, visit https://recovercovid.org/. We gratefully acknowledge the contributions of the RECOVER consortium. We would like to thank the National Community Engagement Group (NCEG), all patients, caregivers, and community representatives, and all the participants enrolled in the RECOVER Initiative. We would like to thank patient representatives Teresa Akintonwa and Lisa McCorkell for their helpful suggestions that have enriched our work. This study was supported by the National Institutes of Health (OT2HL161847-01, Y.C.). Y.C.'s effort has also been supported in part by the National Institutes of Health (1R01LM012607, 1R01AI130460, 1R01AG073435, 1R56AG074604, 1R01LM013519, 1R56AG069880, 1R01AG077820, 1U01TR003709) and the Patient-Centered Outcomes Research Institute (PCORI) Project Program Awards (ME-2019C3-18315 and ME-2018C3-14899). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the PCORI, its Board of

Governors, or the Methodology Committee. Finally, the content is solely the responsibility of the authors and does not necessarily represent the official views of the RECOVER Initiative, the NIH, or other funders.

Author contributions

Authorship was determined using *ICMJE* recommendations. D.Z., B.Z., Q.W., J.S.M., C.B.F., and Y.C. designed methods and experiments; D.J.C., R.J., R.C.K., R.L.R, S.R., D.A.W., C.B.F. provided the datasets for data analysis; D.Z. and B.Z. conducted data analysis; Y.L., D.Z., and B.Z. visualized the analysis results; D.Z., B.Z., J.S.M., C.B.F., and Y.C. interpreted the results; D.Z., B.Z., T.Z., and Y.C. drafted the main manuscript. D.Z., B.Z., Q.W., T.Z., J.T., Y.L., J.C., H.W., D.J.C., R.J., R.C.K., R.L.R, S.R., D.A.W., M.H., L.W., J.S.M., C.B.F., and Y.C. provided critical edits to the early draft and approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests

R.J. is a consultant for AstraZeneca, Seqirus, and Dynavax, and receives an editorial stipend from Elsevier and Pediatric Infectious Diseases Society and royalties from Up To Date/Wolters Kluwer. S.R. reports prior grant support from GSK and Biofire and is a consultant for Sequiris. All other authors have nothing to disclose.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-55273-z.

Correspondence and requests for materials should be addressed to Yong Chen.

Peer review information *Nature Communications* thanks Dominik Ose and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

[1]The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA. [5]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [6]Abigail Wexner Research Institute Nationwide Children's Hospital, Columbus, OH, USA. [7]Division of Infectious Diseases, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, USA. [8]Department of Neurology, New York University Grossman School of Medicine, New York, NY, USA. [9]Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA. [10]Vanderbilt Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN, USA. [11]Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, Aurora, CO, USA. [12]Department of Anesthesiology, University of Michigan, Ann Arbor, MI, USA. [13]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA. [14]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada. [15]Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [16]Leonard Davis Institute of Health Economics, Philadelphia, PA, USA. [17]Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA. [18]Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA. [19]These authors contributed equally: Dazheng Zhang, Bingyu Zhang. ✉e-mail: ychen123@upenn.edu