scientific data

OPEN

DATA DESCRIPTOR

# COVID-19 patient accounts of illness severity, treatments and lasting symptoms

Moriah E. Thomason[1,2,3] ✉, Denise Werchan[1,2] & Cassandra L. Hendrix[1]

**First-person accounts of COVID-19 illness and treatment can complement and enrich data derived from electronic medical or public health records. With patient-reported data, it is uniquely possible to ascertain in-depth contextual information as well as behavioral and emotional responses to illness. The Novel Coronavirus Illness Patient Report (NCIPR) dataset includes complete survey responses from 1,584 confirmed COVID-19 patients ages 18 to 98. NCIPR survey questions address symptoms, medical complications, home and hospital treatments, lasting effects, anxiety about illness, employment impacts, quarantine behaviors, vaccine-related behaviors and effects, and illness of other family/household members. Additional questions address financial security, perceived discrimination, pandemic impacts (relationship, social, stress, sleep), health history, and coping strategies. Detailed patient reports of illness, environment, and psychosocial impact, proximal to timing of infection and considerate of demographic variation, is meaningful for understanding pandemic-related public health from the perspective of those that contracted the disease.**

## Background & Summary

Major discoveries about COVID-19 illness, susceptibility, transmission, and human behavior have been unearthed through utilization of rich medical and public digital record systems. Chasms in health inequity have been revealed[1–3]. Discrete spatiotemporal patterns of public health behavior have been characterized[4,5]. Individual- and community-level risk factors underlying local transmission of COVID-19 have been identified[6,7]. Examination of internet searches has revealed that (i) information flow about COVID-19 is inversely relates to positive cases[8] and (ii) there have been population-level shifts over the course of the pandemic from searches pertaining to activity/fitness to more sedentary activities and dietary supplements[9]. Overall, publicly available large-scale databases are significant sources of information that have been rapidly deployed to identify crucial determinants of health, aspects of transmission, and core human behavioral adaptations in the context of COVID-19.

The challenge, however, is that these studies are limited by the constraints of electronic fields that serve narrow functions and lack nuance that is intrinsic of individual human stories. As a specific example, electronic medical records (EMR) systems include lists of symptoms, physical examination and laboratory results, treatments, diagnoses and basic demographic information. EMR are not intended to capture information about patient perceptions, and yet we know that the subjective experiences and outcomes of one person to the next are variable and also predictive of future health[10,11]. Circumstances of illness occur in diverse socioemotional contexts and relate to other events occurring in an individual's life. The goal of NCIPR was to aggregate a sizable dataset of first-person accounts of COVID-19 illness, risk, and recovery. NCIPR survey data can be used to weave individual strands of history, environment, perspective, and health together to make new discoveries that will compliment and enrich knowledge about COVID-19 that has been derived from medical and public digital record systems. Further, the NCIPR dataset contains measures of lasting secondary effects of COVID-19 infection that are not readily available in medical record systems. In summary, the data available in this data set are different from many other publicly available data sets because these are self-reported patient data in a large sample with notable variability in illness severity, timing, and treatment.

[1]Department of Child and Adolescent Psychiatry, New York University Medical Center, New York, USA. [2]Department of Population Health, New York University Medical Center, New York, NY, USA. [3]Neuroscience Institute, NYU Langone Health, New York, NY, USA. ✉e-mail: moriah.thomason@nyulangone.org

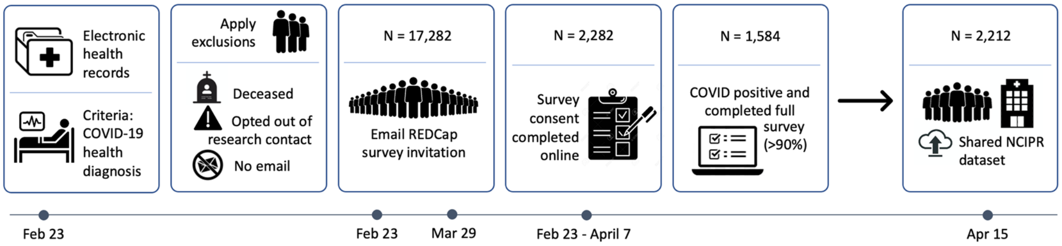
Electronic health records
Criteria: COVID-19 health diagnosis
Apply exclusions
Deceased
Opted out of research contact
No email
N = 17,282
Email REDCap survey invitation
N = 2,282
Survey consent completed online
N = 1,584
COVID positive and completed full survey (>90%)
N = 2,212
Shared NCIPR dataset
Feb 23
Feb 23
Mar 29
Feb 23 - April 7
Apr 15

Fig. 1 A schematic overview of the study design and data collection workflow.

| Domain | Topics addressed |
|---|---|
| COVID-19 illness | Date if illness; symptoms experienced; exposure; length of illness; fever; perceived severity |
| COVID-19 treatments | Hospitalization; ICU admission; medical treatments; at-home treatments; medications; imaging |
| COVID-19 testing | Timing; type; facility; presence of symptoms |
| COVID-19 impacts | Loss of income; time off work; quarantine behaviors; perceived life disruption; change in social support, sleep, energy levels, stress, relationship satisfaction |
| COVID-19 perceptions | Anxiety about illness; satisfaction with medical care; opinion about when things will go back to normal |
| COVID-19 lasting effects | Length of symptoms; types of symptoms; specific mood disturbances; specific cognitive disturbances; estimate of time to full return to health |
| Physical characteristics | Age; height; weight; blood type |
| Health characteristics | Preexisting health conditions; prior substance use and mental health treatment; current stress level; prior tonsillectomy |
| Vaccine information and attitudes | Date received; manufacturer; side effects; if not vaccinated, attitudes about vaccination for self and/or children; plans to relax COVID-19 safety behaviors after vaccination; if breastfeeding, attitudes about vaccination and infant side effects |
| Demographic and financial context | Gender identity; employment (self and partner); income change due to COVID-19; stability of housing; public assistance; medical insurance; satisfaction with financial situation; MacArthur Scale of subjective social status |
| Home environment | Pets; number of individuals living in home; number of household members that became ill; number of bedrooms in home |
| Patient behavior | COVID safety behavior; coping strategies; drug, nicotine and alcohol consumption; exercise; use of meditation/mindfulness; religious practices; family/friend support; screen use; social media use |
| Perceived discrimination | Amount; kind; distress |

Table 1. Summary of measurement domains assessed by the NCIPR Survey. Data across these domains is contained within the New York NCIPR dataset. The NCIPR questionnaire also includes questions about child ages, breastfeeding, education, race/ethnicity, income, number of bedrooms in home, utilization of public assistance, and preferred medical health system. For data release and compliance with regulation on indirect identifiers and patient confidentiality, these are removed from released data, as described below.

The Novel Coronavirus Illness Patient Report (NCIPR) survey was developed in November 2020 and published in the U.S. National Library of Medicine (NLM) Disaster Management Resources (id:24224) as well as the Open Science Framework (OSF; https://doi.org/10.17605/OSF.IO/82RKJ)[12] in early March 2021. Patients with COVID-19 diagnoses were identified from within the New York University Langone Medical Center (NYU Langone) EMR system. This EMR data extraction occurred on February 23, 2021. The full workflow is depicted in Fig. 1, including record extraction, invitation, online consent, and resulting dataset. Two waves of recruitment invitations were implemented, occurring on February 23 and March 29, 2021. Between waves, four new questions were added to gather additional data on lasting symptom complaints, including duration of symptoms, categories of mood symptoms, and two questions about lasting cognitive complaints. Additionally, five questions were added about blood type, height and weight, history of tonsillectomy and the Macarthur Ladder[13]. The survey was closed to potential respondents on April 7, 2021. Curated, notated data were uploaded to OSF April 15, 2021, and data revised based on external input were uploaded to OSF on September 28, 2021.

The primary goal motivating collection of the New York NCIPR dataset was to obtain a record of the subjective experiences of those ill with COVID-19, proximal to the time of illness. Along with this, we asked targeted questions that could address topics such as unexpected side effects (e.g., hair loss), lasting illness sequalae, vaccine hesitancy, and potential areas of underlying vulnerability. As a result, the NCIPR dataset can be used to address a large number of questions that remain unanswered about COVID illness, about human behavior, and about environmental determinants of health. Rapid placement of the data in the public domain better assures that investigation of these and other topics will commence quickly and will be rapidly communicated to wide audiences.

## Methods

### Survey design.

The NCIPR survey was developed to assess COVID-19 symptoms, medical complications, home and hospital treatments, lasting effects, anxiety about illness, employment impacts, quarantine behaviors, vaccine-related behaviors and effects, and illness of other family/household members. The NCIPR also includes

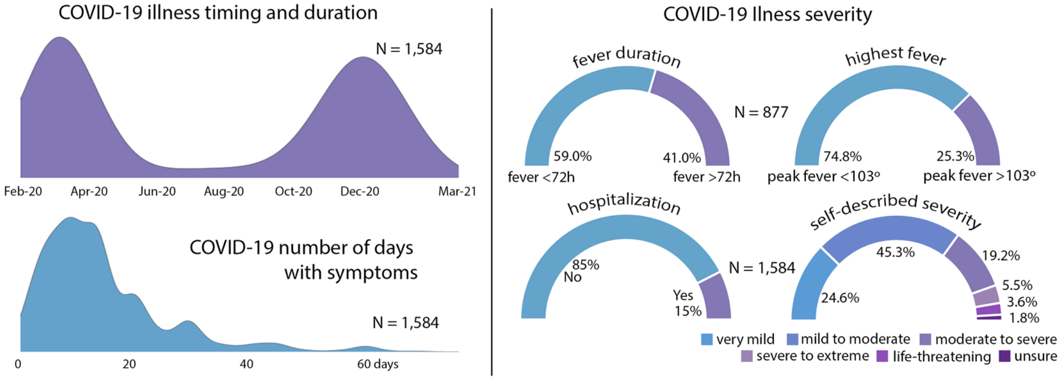

**Fig. 2** Overview of illness severity in the N = 1,584 COVID-19 quality validated sample. Fever peak and length are given only for those that endorsed having had a fever while infected (N = 877).

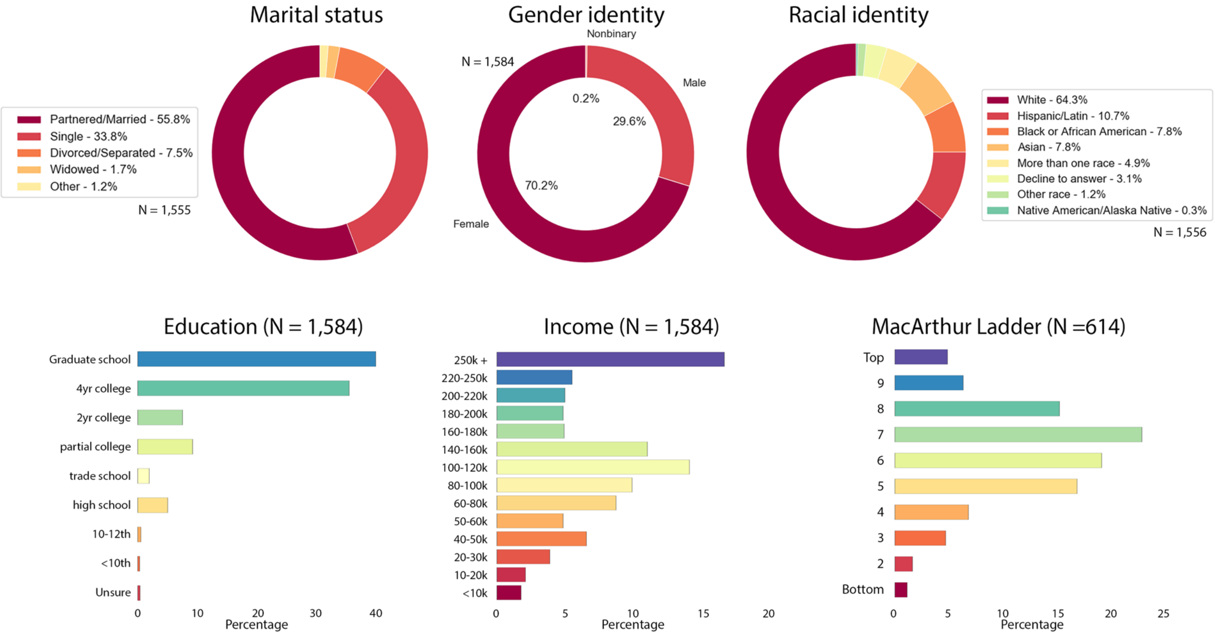

**Fig. 3** COVID-19 infected sample demographics N = 1,584. MacArthur Ladder responses are only available for those that responded after March 29, 2021 (N = 614), as this question was added between the two recruitment invitations.

questions that address age, financial security, perceived discrimination, pandemic impacts (relationship, social, stress, sleep), health history, and behavioral coping strategies. A subset of questions were adapted from established Common Data Elements for mental health, specifically, the NLM Disaster Management Resources COVID-19 and Perinatal Experiences (COPE) questionnaire, https://doi.org/10.17605/OSF.IO/UQHCV; the Williams Perceived Discrimination Scale[14]; and the Fletcher measure of Perceived Relationship Quality[15]. Table 1 provides a summary of domains covered by the full NCIPR survey.

**Ethical approval.** The research protocol for this study was approved by the NYU Langone Institutional Review Board (IRB). Only patients that had previously consented to be contacted about research opportunities were eligible for invitation into the study. Participants provided consent to share de-identified survey data. The approved study protocol included sharing of de-identified data with outside researchers or research databases.

**Recruitment and survey administration.** A search of the NYU Langone Health record system identified all individuals ages 18 and older that had been diagnosed with COVID-19 based on symptoms or lab results. Individuals (1) with email contact, (2) not deceased, and (3) not designated as having previously opted out of research contact were eligible to participate. After application of these exclusions, 17,282 individuals were sent an email inviting them to participate in a 10 to 15-minute survey. Compensation was entry into an end of week drawing for a $25 Amazon gift card. Study data were collected and managed using REDCap electronic data capture tools hosted at NYU Langone University[16,17]. The measure was administered in English. Survey questions

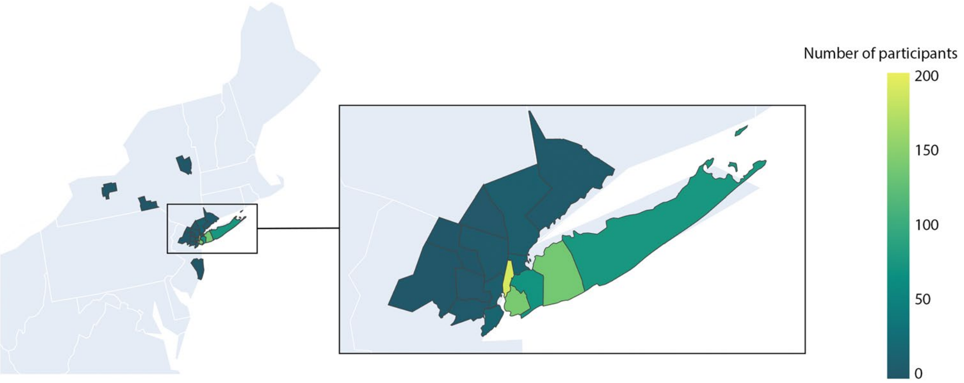

**Fig. 4** Geographical location of COVID-19 patient survey respondents. Geographical information was only available for those that provided zip code data (N = 697).

| Variable name | Variable definition |
|---|---|
| which_NCIPR | (1) NCIPR wave 1 February 23, 2021; (2) NCIPR wave 2 March 31, 2021 |
| complete_binary | (0) incomplete (n = 483); (1) complete (n = 1,729) |
| why_incomplete | (1) complete (n = 1,729); (2) survey administration error (n = 338); (3) incomplete survey (n = 145) |
| covid_self_report | (0) report no prior COVID-19 illness (n = 65); (1) confirm COVID-19 prior illness (n = 2,147) |
| DOB_age_out_of_range | (0) date of birth age = 18–100 years (n = 2,157); (1) date of birth age = <18 or age >100 (n = 55) |
| COVID_date_out_of_range | (0) Feb 2020 - March 2021 (n = 2,192); (1) dates in range not selected (n = 20) |
| quality_check_flag | (0) none (n = 1,857); (1) ≥1 implausible response (e.g., 6'20" tall) (n = 4); (2) ≥1 inconsistent response (What is your current age? [db_52] ≠ reported date of birth +/− one year) (n = 68); (3) inconclusive (e.g., age or DOB response not provided) (n = 283) |
| data_correction | (0) no correction; (1) typo in age or height; original data unchanged but [quality_check_flag] changed to '0' (n = 7) |
| excluded_sample | (0) included (n = 1,584); exclusions filtered in the following order: (1) incomplete (n = 145); (2) survey admin error (n = 338); (3) [covid_self_report] = '0' (n = 65); (4) DOB provided out of range (n = 46); (5) [quality_check_flag] = '1' or '2' (n = 19); (6) COVID-19 illness date inconclusive (n = 15) |
| age_calculated | Participant reported date of birth [db_2] converted to age in years |

**Table 2.** Summary of variables added to dataset during preparation and validation steps.

included questions about whether the individual believed themself to have had COVID-19, whether they had a positive antibody test, whether they had a positive PCR test, where they were tested and how mild to severe they rate their illness. Because participants were invited on the basis of a COVID-19 code in the medical record system and because testing asymptomatic individuals on the basis of exposure was the predominant standard of care in New York City in this time frame, asymptomatic cases that tested positive may be discoverable in the data set.

**Sample description.** The NCIPR dataset contains data from 2,212 individual respondents. 2,147 of these respondents confirm having been ill with COVID-19 in addition to having COVID-19 diagnosis in their medical record. However, description of illness severity and demographics provided here are restricted to 1,584 cases that passed the Technical Validation steps described in the section below. Timing of COVID-19 illness in the sample reflects peak prevalence rates in March 2020 and January 2021 (Fig. 2). Illness severity varied across the sample, as seen in length of illness, fever duration, peak fever, hospitalizations, and in self-reported illness severity ratings (Fig. 2). Sample demographic data are provided in Fig. 3. Respondent ages range from 18 to 98 years old. Due to a survey administration error described below, complete data are available at a ratio of ~2:1, females to males.

**Geo-positioning of COVID-19 survey respondents.** Geographical information about survey respondents was derived from a subset of patients (N = 697) that provided consent to future contact within the online consent form. Those that made this selection were asked to provide contact information and zip code data. Zip codes were converted to corresponding Federal Information Processing System (FIPS) codes. The distribution of patient FIPS is displayed in Fig. 4. The majority reside in Manhattan, Brooklyn and Long Island. A small number provided zip codes in states other than New York, New Jersey and Connecticut, N = 9. Geographic restriction of survey data limits the generalizability of these data to other parts of the United States and world.

## Data Records

The dataset resulting from the NCIPR survey is stored in a CSV format via the *Open Science Framework* open access platform at https://doi.org/10.17605/OSF.IO/82RKJ[12]. Each row represents one respondent and each column represents a variable. The file includes every survey respondent except for those who completed the

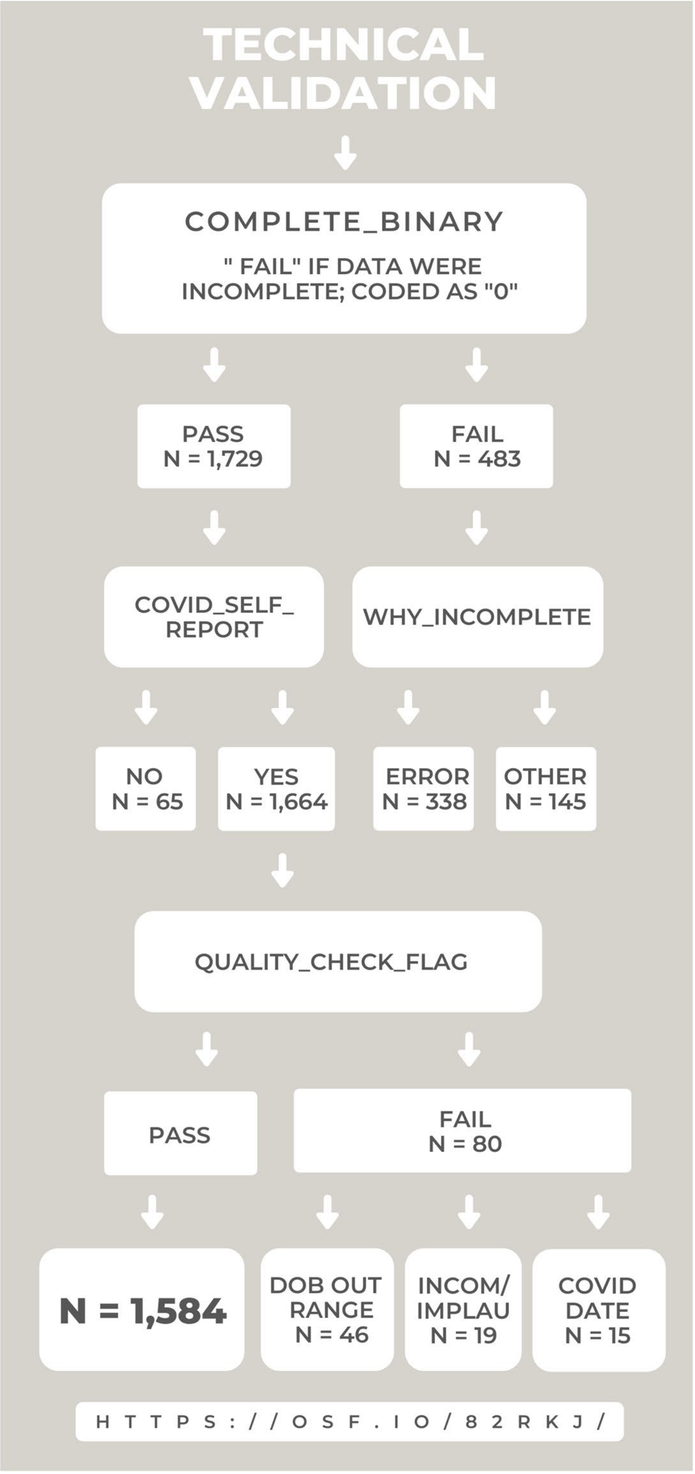

**Fig. 5** Overview of technical validation workflow and number of cases excluded at each step. 2,212 cases are included on the data release. 1,584 are coded as having passed all technical validation criteria. Abbreviations: date of birth, DOB; incompatible, INCOMP; implausible, IMPLAU. Data available via Open Science Framework (OSF).

consent form only (N = 68). Date of birth was converted to age in years, variable name [age_calculated]. A second variable, [db_52], is the age in years provided by the participant. Inclusion of both age-related data fields was intentional, as this provides a means of data validation, described in more detail below. Ordering of the variables in the CSV files reflects the order in which items were administered. During data preparation and validation, 10 variables were added to aid in future data processing. Table 2 summarizes the variables that were added to the raw data set during quality assessment and data validation. Please note that the dataset includes variable, "which_ncipr" [1,2] and variable, "todays_date", that are indicators of respondent involvement in first or second wave and date completed, respectively. Participants that identified as male or as non-binary gender were invited again in wave two, as this provided opportunity for them to complete the survey. All participants that responded more than once are designated with the same subject ID. We selected to filter out incomplete wave 1 data in selection of the 1,584 included cases, and as such, if applied, the "excluded_sample" filter, would effectively remove all duplicate respondents.

Additional files released with the primary dataset (.csv) are: (a) the NCIPR questionnaire (.pdf), (b) the NCIPR demographics form (.pdf), (c) the REDCap instrument files (.zip), and (d) the variable definition file (.csv). All are accessible via the Open Science Framework (OSF) open access platform. The questionnaires include response options for each question along with the coding used for each variable. The REDCap files shared via https://doi.org/10.17605/OSF.IO/82RKJ include the nine additional questions added between recruitment waves. As is evident in ordering of variables in the CSV file, the NCIPR covid illness survey was administered prior to the NCIPR demographics survey.

**Survey administration error.** An error in branching logic was identified after the first wave of data collection, such that respondents who did not endorse being female were not offered the majority of questions about COVID illness. This error was identified within the first 24 hour of survey administration and was corrected. This error resulted in a systematic loss of data in 322 male and self-describing gendered participants for COVID illness questions. Wave-one male and self-describing gendered cases are included in the shared dataset and are designated as such, as referenced in Table 2 and in the added variable [why_incomplete] = 2. One repercussion of this error is that the ratio of females to males is higher for wave one collection. As mentioned above, cases responding to wave one or wave two are designated by variable "which_ncipr", making it possible to take this into account during analyses.

**Removal of indirect identifiers.** Confidentiality and anonymity are key ethical considerations when publishing or sharing data relating to individuals[18]. Indirect identifiers removed from the dataset are indicated in the data variable definition file available on OSF, https://doi.org/10.17605/OSF.IO/82RKJ. Indirect identifiers removed include race, ethnicity, income, education, DOB, ages of children, number of bedrooms in home, breastfeeding questions, use of public assistance, number of adults and children in home, and affiliation with NYU hospital system. Further, all dates in the dataset were converted to Month-Year format (e.g. Mar-21) and individuals age 90 or older were edited to 89+ to disallow potential re-identification.

## Technical Validation

Data assurance and quality checking were performed using R version 4.0.2 and Excel. Table 2 provides a summary of variables added to the dataset during quality validation steps, inclusive of QA/QC codes assigned to survey respondents. Criterion assessed for determinations about quality of patient responses included isolating implausible and/or inconsistent responses. Patients were flagged [quality_check_flag] as (1) "implausible" if they provided a height feet value greater than 7, or a height inches value greater than 12; (2) "inconsistent" if the self-reported date of birth (DOB) and current age were incongruent (defined as different by >1 year); or (3) "inconclusive" if DOB or age in years was not provided. It was noted that 5 individual respondents gave their full height in inches (e.g., 5.2 was entered as feet and 62 was entered as inches), and 2 participants typed a decimal point before self-reported age in years that matched the date of birth provided (e.g., born in 1997 and provided age 0.24). For those 7 cases, the [quality_check_flag = 1] was changed to [quality_check_flag = 0] and they were included in the final sample, [final_sample = 1], but the raw data causing the flag was not changed. Patient age was computed based on DOB and inserted as a new variable in the dataset [age_calculated]. Findings from these preparation and validation steps guided selection of a final sample that is coded as [excluded_sample] = '0' in the released data; these are the 1,584 described above as passing technical validation for which group level demographics are provided. The number of cases excluded at each step of QA/QC is depicted in Fig. 5.

## Code availability

No new code was used or developed for the study.

Received: 7 May 2021; Accepted: 4 November 2021;
Published online: 10 January 2022

## References

1. Azar, K. M. J. *et al*. Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California. *Health affairs (Project Hope)* **39**, 1253–1262 (2020).
2. Adhikari, S. *et al*. Assessment of Community-Level Disparities in Coronavirus Disease 2019 (COVID-19) Infections and Deaths in Large US Metropolitan Areas. *JAMA network open* **3**, e2016938 (2020).
3. Krieger, N., Waterman, P. D. & Chen, J. T. COVID-19 and Overall Mortality Inequities in the Surge in Death Rates by Zip Code Characteristics: Massachusetts, January 1 to May 19, 2020. *American journal of public health* **110**, 1850–1852 (2020).
4. Borgonovi, F. & Andrieu, E. Bowling together by bowling alone: Social capital and COVID-19. *Soc Sci Med* **265**, 113501 (2020).
5. Fu, X. & Zhai, W. Examining the spatial and temporal relationship between social vulnerability and stay-at-home behaviors in New York City during the COVID-19 pandemic. *Sustainable cities and society* **67**, 102757 (2021).
6. Sugg, M. M. *et al*. Mapping community-level determinants of COVID-19 transmission in nursing homes: A multi-scale approach. *The Science of the total environment* **752**, 141946 (2021).
7. Andersen, L. M., Harden, S. R., Sugg, M. M. P., Runkle, J. D. P. & Lundquist, T. E. Analyzing the spatial determinants of local Covid-19 transmission in the United States. *The Science of the total environment* **754**, 142396 (2021).
8. Chundakkadan, R. & Ravindran, R. Information flow and COVID-19 recovery. *World development* **136**, 105112 (2020).

9. Mayasari, N. R. *et al*. Impacts of the COVID-19 Pandemic on Food Security and Diet-Related Lifestyle Behaviors: An Analytical Study of Google Trends-Based Query Volumes. *Nutrients* **12** (2020).
10. Danese, A. & Widom, C. S. Objective and subjective experiences of child maltreatment and their relationships with psychopathology. *Nature human behaviour* **4**, 811–818 (2020).
11. Hertzman, C. & Boyce, T. How experience gets under the skin to create gradients in developmental health. *Annu Rev Public Health* **31**, 329–347 323p following 347 (2010).
12. Thomason, M. E., Werchan, D. & Hendrix, C. L. Novel Coronavirus (COVID) Illness – Patient Report (NCIPR). *Open Scinece Framework* https://doi.org/10.17605/OSF.IO/82RKJ (2021).
13. Adler, N. E., Epel, E. S., Castellazzo, G. & Ickovics, J. R. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy white women. *Health psychology: official journal of the Division of Health Psychology, American Psychological Association* **19**, 586–592 (2000).
14. Williams, D. R., Yan, Y., Jackson, J. S. & Anderson, N. B. Racial Differences in Physical and Mental Health: Socio-economic Status, Stress and Discrimination. *Journal of health psychology* **2**, 335–351 (1997).
15. Fletcher, G. J. O., Simpson, J. A. & Thomas, G. The Measurement of Perceived Relationship Quality Components: A Confirmatory Factor Analytic Approach. *Personality and Social Psychology Bulletin* **26**, 340–354 (2000).
16. Harris, P. A. *et al*. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics* **42**, 377–381 (2009).
17. Harris, P. A. *et al*. The REDCap consortium: Building an international community of software platform partners. *Journal of biomedical informatics* **95**, 103208 (2019).
18. Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J. & Altman, D. G. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Bmj* **340**, c181 (2010).

## Acknowledgements

This project was supported by R01MH126468, R01MH125870, and a COVID-19 supplement award to MET from the National Institutes of Health connected to R34DA050287. The authors thank Autumn Austin, Carly Lenniger, and Amin Majbri for contributions to recruitment and data collection. The authors thank participants who generously shared their time and who expressed interest in helping researchers and clinicians better understand the varied experiences and circumstances surrounding COVID-19 illness and recovery.

## Author contributions

M.T. designed the experiments. D.W. and C.H. analyzed the data. M.T. contributed materials. M.T., D.W. and C.H. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.E.T.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2022