Estiri et al. BMC Medicine (2021) 19:249
https://doi.org/10.1186/s12916-021-02115-0

BMC Medicine

RESEARCH ARTICLE Open Access

Evolving phenotypes of non-hospitalized patients that indicate long COVID

Hossein Estiri[1,2*], Zachary H. Strasser[1,2,3], Gabriel A. Brat[3], Yevgeniy R. Semenov[4], The Consortium for Characterization of COVID-19 by EHR (4CE), Chirag J. Patel[3] and Shawn N. Murphy[1,5,6]

Abstract

Background: For some SARS-CoV-2 survivors, recovery from the acute phase of the infection has been grueling with lingering effects. Many of the symptoms characterized as the post-acute sequelae of COVID-19 (PASC) could have multiple causes or are similarly seen in non-COVID patients. Accurate identification of PASC phenotypes will be important to guide future research and help the healthcare system focus its efforts and resources on adequately controlled age- and gender-specific sequelae of a COVID-19 infection.

Methods: In this retrospective electronic health record (EHR) cohort study, we applied a computational framework for knowledge discovery from clinical data, MLHO, to identify phenotypes that positively associate with a past positive reverse transcription-polymerase chain reaction (RT-PCR) test for COVID-19. We evaluated the post-test phenotypes in two temporal windows at 3–6 and 6–9 months after the test and by age and gender. Data from longitudinal diagnosis records stored in EHRs from Mass General Brigham in the Boston Metropolitan Area was used for the analyses. Statistical analyses were performed on data from March 2020 to June 2021. Study participants included over 96 thousand patients who had tested positive or negative for COVID-19 and were not hospitalized.

Results: We identified 33 phenotypes among different age/gender cohorts or time windows that were positively associated with past SARS-CoV-2 infection. All identified phenotypes were newly recorded in patients' medical records 2 months or longer after a COVID-19 RT-PCR test in non-hospitalized patients regardless of the test result. Among these phenotypes, a new diagnosis record for anosmia and dysgeusia (OR 2.60, 95% CI [1.94–3.46]), alopecia (OR 3.09, 95% CI [2.53–3.76]), chest pain (OR 1.27, 95% CI [1.09–1.48]), chronic fatigue syndrome (OR 2.60, 95% CI [1.22–2.10]), shortness of breath (OR 1.41, 95% CI [1.22–1.64]), pneumonia (OR 1.66, 95% CI [1.28–2.16]), and type 2 diabetes mellitus (OR 1.41, 95% CI [1.22–1.64]) is one of the most significant indicators of a past COVID-19 infection. Additionally, more new phenotypes were found with increased confidence among the cohorts who were younger than 65.

Conclusions: The findings of this study confirm many of the post-COVID-19 symptoms and suggest that a variety of new diagnoses, including new diabetes mellitus and neurological disorder diagnoses, are more common among those with a history of COVID-19 than those without the infection. Additionally, more than 63% of PASC phenotypes were observed in patients under 65 years of age, pointing out the importance of vaccination to minimize the risk of debilitating post-acute sequelae of COVID-19 among younger adults.

Keywords: Post-acute sequelae of SARS-CoV-2, Electronic health records, Phenotypes, Machine learning

* Correspondence: hestiri@mgh.harvard.edu
[1]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA 02114, USA
[2]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
Full list of author information is available at the end of the article

Background

The onslaught of the COVID-19 pandemic in the USA and around the world was relentless. For many, recovery from the acute phase of the SARS-CoV-2 infection, the coronavirus that causes COVID-19, may be grueling with a debilitating second act. A collection of persistent physical (e.g., fatigue, dyspnea, chest pain, cough), psychological (e.g., anxiety, depression, post-traumatic stress disorder), and neurocognitive (e.g., impaired memory and concentration) symptoms can appear and last for weeks or months in patients after acute COVID-19 [1–8]. Many of the symptoms characterized as the post-acute sequelae of COVID-19 (PASC) could have multiple causes.

So far, a number of studies have been published on PASC [1–7, 9, 10], but most have small samples and case series or rely on self-reports. Carfi et al. assessed 179 hospitalized COVID patients in Italy at an average of 60 days after the onset of symptoms using a standard questionnaire [11]. Only 12.6% were completely free of all COVID-19 symptoms, and 55% had 3 or more symptoms. The most common symptoms were fatigue, dyspnea, joint pain, and chest pain. Chopra et al. performed an observational study of 488 patients who were hospitalized 60 days after their discharge with a phone survey [12]. The most common persistent symptoms were cough, dyspnea, persistent loss of taste or smell, and worsening difficulty completing activities of daily living. Huang et al. performed one of the larger cohort studies where they analyzed 1733 COVID patients discharged from a hospital in China with a questionnaire at 6 months [13]. They identified fatigue, muscle weakness, sleep difficulties, anxiety, and depression as the most common symptoms 6 months after the initial diagnosis.

These studies are all case series, focusing only on patients with COVID-19. Additionally, prior PASC studies often focus on patients with severe COVID-19 symptoms after hospitalization. It is unclear whether the identified persistent symptoms hold true among COVID patients not hospitalized. Furthermore, many of the published studies are based on small cohorts (several hundred COVID-19 patients were analyzed) and relied on self-reported outcomes which can embody potential biases due to, for example, exaggeration of symptoms [14].

There have also been a number of less commonly reported symptoms including ocular inflammation [15], cardiac involvement [16, 17], autonomic instability [18], recurrent *Pseudomonas* infections [19], persistent mucous secretion [20], micro-structural changes to the brain [21], and Guillain-Barre syndrome [22]. A large cohort analyzing the ICD-10 (the 10th Revision of the International Statistical Classification of Diseases and Related Health Problems) diagnoses in the electronic health record between patients with and without a history of COVID could help clarify the actual association with the disease.

We present the results from a retrospective cohort study of over 97,000 patients with an RT-PCR test for COVID-19 in a Mass General Brigham (MGB) facility. We detected de novo phenotypes that appeared for the first time in EHRs at two temporal windows of 3–6 and 6–9 months after a COVID-19 test for both COVID-positive and COVID-negative patients. Leveraging MLHO, a computational framework developed for knowledge discovery from electronic health records (EHRs) [23–25] with a validated utility for studying and modeling post-COVID outcomes [26, 27] augmented with clinical expertise, we identified 33 phenotypes in different age/gender groups or time windows positively associated with a recent/past SARS-CoV-2 infection. All identified phenotypes were newly recorded in the patients' medical records 2 months or longer after a COVID-19 RT-PCR test in non-hospitalized patients regardless of the test result.

Methods

We utilized longitudinal EHR diagnosis records from all patients who tested for SARS-CoV-2 infection—reverse transcription polymerase chain reaction (RT-PCR)—between March 2020 and June 2021 in a Mass General Brigham (MGB) facility. We limited the patient cohort to those who were alive and not hospitalized. To increase the confidence that a patient in our cohort would likely seek care within MGB in the post-COVID era, we further narrowed the study population to patients who had two diagnosis records, 6 months apart, in our electronic data repositories since 2010. We also excluded patients who had a diagnosis code referring to past COVID-19 but having a negative RT-PCR test in the MGB records due to our inability to approximate the infection date. The use of clinical data in this study was approved by the MGB Institutional Review Board with a waiver of informed consent.

Phenotype coding

To construct the feature space, we utilized EHR diagnoses recorded in the ICD-9 and ICD-10 codes (the 9th and 10th Revisions of the International Statistical Classification of Diseases and Related Health Problems). To represent the phenotypes for the analyses, we mapped the ICD-9/10 diagnosis codes to a unique phenotype code (PheCode) from the phenome-wide association studies (PheWAS) [28, 29] groups of phenotypes. We assigned a temporal buffer of 2 months after the RT-PCR test as a proxy for the acute phase in COVID-19 patients and used the first observation of phenotypes that were recorded for the first time after the acute phase (Fig. 1). Using this temporal segmentation, we further limited the data, by only using the first observation of the records (to minimize the problem list repetitions) and only considered the diagnosis records that for the

first time appeared in a patient's medical records 2 months or longer after the RT-PCR test—see Additional file 1: Fig. S1. As such, the feature space contained all PheCodes that were recorded for the first time in a patient's longitudinal EHR data 2 months or later after the COVID-19 RT-PCR test, regardless of the test result.

MLHO framework

To robustly identify the phenotypes that are positively associated with a recent positive test for COVID-19, we applied a multivariate temporal approach to classify past RT-PCR test results from the post-test clinical data. The classification algorithm here is not intended for the purpose of classification. Rather, we performed "postdiction," which is the "assertion or deduction about something in the past," [30] aiming to identify the features (i.e., phenotype) that carry information to make such an assertion about the past event. To do so, we leveraged the MLHO framework [26], which includes a suite of computational algorithms [23, 26] specifically designed for modeling and phenotyping clinical data. We followed a similar analytic process used by Estiri et al. [31] that was used to identify the risk factors for COVID-19 mortality from EHR data. From the MLHO framework, the computational process involved applying the minimize sparsity, maximize relevance (MSMR) algorithm [23, 32, 33]; clinical expertise; and multivariate boosting logistic regression, to compute a composite confidence score for identifying the phenotypes that are positively associated with a past RT-PCR test (see MLHO phenotype selection criteria in the additional file).

All analyses were conducted in R statistical language.

Cohort stratification

To increase specificity, we stratified the analyses by age and gender in a nested structure. This resulted in the following strata: (1) all patients, (2) 65 and older, (3) under 65, (4) 65 and older female, (5) 65 and older male, (6) under 65 female, and (7) under 65 male. In addition to stratifying the cohort, we controlled for the age and gender (in gender-agnostic models) of the patient. For the phenotypes identified by MLHO in each stratified model, we trained standard generalized logistic regression models controlling for age and gender and extracted multivariate odds ratios (ORs) along with *p*-value (Wald's test) and 95% confidence intervals using a profiled log-likelihood.

Clinical validation via chart reviews

Due to the known reliability issues of EHR diagnosis records [33, 34], we validated the phenotypes identified by MLHO through chart reviews. A clinical expert reviewed the clinical notes and longitudinal records for a random sample of five patients for each phenotype identified by MLHO with an 80-plus confidence score. The chart review required reviewing the clinical notes at the time of the diagnostic code to determine whether the phenotype was actually present at the encounter and whether this was a new symptom or diagnosis since the time of the COVID encounter. If at least three of the randomly sampled five charts verified the phenotype's presence and its recent appearance or diagnosis, then the phenotype was included in the final analysis.

Results

From over 397,000 patients who tested for COVID-19 in an MGB facility with a nasal swab, 210,949 met our inclusion/exclusion criteria, including 52,491 patients with positive test results. After applying the approach for keeping records, 96,025 patients remained in our final study cohort, which means 45.71% of the outpatient cohort who tested for the infection at an MGB facility had a new phenotype record in their EHRs 2 months or longer after the RT-PCR test. A total of 22,475 (23.41%) of these patients were positive for the SARS-CoV-2 virus (Additional file 1: Fig. S2 and Table S1). After the sparsity screening (i.e., removing low-prevalence [< 0.22%] phenotypes from sub-cohorts), 354 and 334 phenotypes were evaluated in the full cohorts during the 3–6- and 6–9-month temporal windows.

Overall, MLHO identified 41 phenotypes in different age/gender groups and/or time windows as positively associated with a past positive COVID-19 test, with a MLHO confidence score higher than 80. All identified phenotypes were newly recorded in the patients' medical records 2 months or longer after a COVID-19 RT-PCR test in non-hospitalized patients regardless of the test

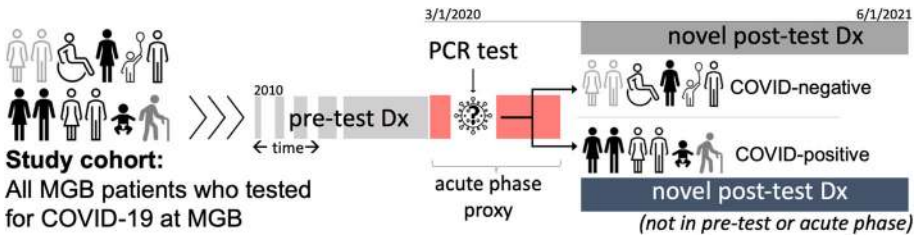

Fig. 1 Study cohort, temporal segmentation, and diagnoses recording. The study evaluated diagnosis records from the post-RT-PCR test from all patients who tested for COVID-19 at an MGB facility and had reliable EHR data longevity

result. We performed chart reviews on 215 randomly sampled patients to validate MLHO's findings. For nearly all of the phenotypes, the details and descriptions provided in the clinical notes matched with the assigned phenotype for that chart (Additional file 1: Table S2). For 33 of the phenotypes (Figs. 2 and 3), the majority of the random samples of notes reviewed were suggestive that the phenotype was new since the time of COVID. Accordingly, we removed 8 phenotypes due to the likelihood they were present pre-COVID based on the notes, despite the use of a new ICD-9/10 record since the COVID-19 diagnosis. For the 33 phenotypes, multivariate odds ratios (ORs), 95% confidence intervals, and MLHO's confidence scores (CSs) are provided below—also available in Additional file 1: Table S3.

The results demonstrated extremely high confidence (> 97%) in eleven phenotypes, which in the overall cohort and/or one or more sub-cohorts associate with a positive past COVID-19 infection. Seven were very high among the entire population in the 3–6-month window. Alopecia was identified in all iterations of MLHO between months 3 and 6, in the overall cohort (OR 3.09, 95% CI [2.53–3.76], CS 100). It was also specifically seen in those younger and older than 65 cohorts and specifically in women both under and over 65. Similarly, a new diagnosis record of non-specific chest pain was indicative of past COVID-19 infection in the 3–6-month temporal window (OR 1.27, 95% CI [1.09–1.48], CS 100) and particularly among people under 65 (OR 1.30, 95% CI [1.08–1.55], CS 100). Anosmia and dysgeusia were identified in 100% of the MLHO iterations, in the 3–6-month window (OR 2.60, 95% CI [1.94–3.46], CS 100) and continued to be important in the 6–9-month window (OR 2.10, 95% CI [1.40–3.11], CS 100). The phenotype was indicative of past positive COVID-19 in those under 65 and women under 65.

Among other identified phenotypes with 97 and higher confidence scores, chronic fatigue syndrome was seen in both the 3–6-month window (OR 2.60, 95% CI [1.22–2.10], CS 98) and the 6–9-month window (OR 2.03, 95% CI [1.31–3.11]), appearing more prominent in the patients less than 65 and women less than 65. Pneumonia, in the 3–6-month window, had a high confidence score among the overall population (OR 1.66, 95% CI [1.28–2.16], CS 99) and those older than 65 (OR 1.92, 95% CI [1.03–3.46], CS 99). Shortness of breath had high confidence scores in both the 3–6-month window (OR 1.41, 95% CI [1.22–1.64], CS 100) and the 6–9-month window (OR 1.45, 95% CI [1.09–1.93], CS 96). It also was identified as having a high confidence score among those under 65. Finally, palpitations (OR 1.41, 95% CI [1.22–1.64]) type 2 diabetes mellitus (OR 1.41, 95% CI [1.22–1.64]) also had high confidence scores both in the 3-6-month window.

Several phenotypes had very high scores but only within certain time frames and in certain sub-cohorts, for example, iron deficiency anemia in the 6–9-month range for those under 65 (OR 2.02, 95% CI [1.37–2.95], CS 100) and women under 65 (OR 2.10, 95% CI [1.40–3.15], CS 100). Men under 65 were identified with proteinuria (OR 3.19, 95% CI [1.72–5.96], CS 100) in the 3–6-month range and syncope and collapse (OR 4.80, 95% CI [1.56–13.39], CS 99) in the 6–9-month range.

Among other COVID-19-related phenotypes identified as indicators of past COVID-19 infection with a 90 to 96 confidence score were a number of sub-groups. In the 3–6-month window, this includes anemia during pregnancy in women under 65, chronic kidney disease in the cohort older than 65 and women over 65, heart failure with preserved ejection fraction in the cohort older than 65, irregular menstrual cycle in women under 65, neurological disorders in those under 65, and rash and other non-specific skin eruptions in men under 65. In the 6–9-month range phenotypes, with a confidence score in the 90 to 96 window, this includes anemia of chronic disease in women 65 and older, disorders of the conjunctiva in men under 65, dizziness and lightheadedness in women older than 65, irregular menstrual cycle in the total cohort, sensorineural hearing loss in women greater than 65, and vascular dementias for those older than 65 and women older than 65.

Discussion

We identified 33 phenotypes that were indicative of long COVID among non-hospitalized COVID-19 patients. Phenotypes such as alopecia, anosmia, fatigue, shortness of breath, and chest pain have been well documented as common signs and symptoms of PASC [7, 35, 36]. This study shows that these phenotypes are some of the earliest associations with the syndrome seen in the 3–6-month window after the initial infection and some of the most important features for indicating previous COVID-19 infection. All five of these phenotypes (alopecia, anosmia and dysgeusia, shortness of breath, chronic fatigue syndrome, and non-specific chest pain) were documented with high confidence in the 3–6-month window. And while alopecia and non-specific chest pain were not found with high confidence in the 6–9-month window, anosmia and chronic fatigue syndrome continued to be important phenotypes seen in both time periods. Additionally, several phenotypes were identified with similarly high confidence including type II diabetes, pneumonia, proteinuria, and syncope and collapse.

Interestingly, those aged less than 65 had more new phenotypes identified with greater confidence than the cohorts who were older than 65. Over 63% of the identified long COVID phenotypes were observed in past COVID-19 patients who were under 65 years old. These

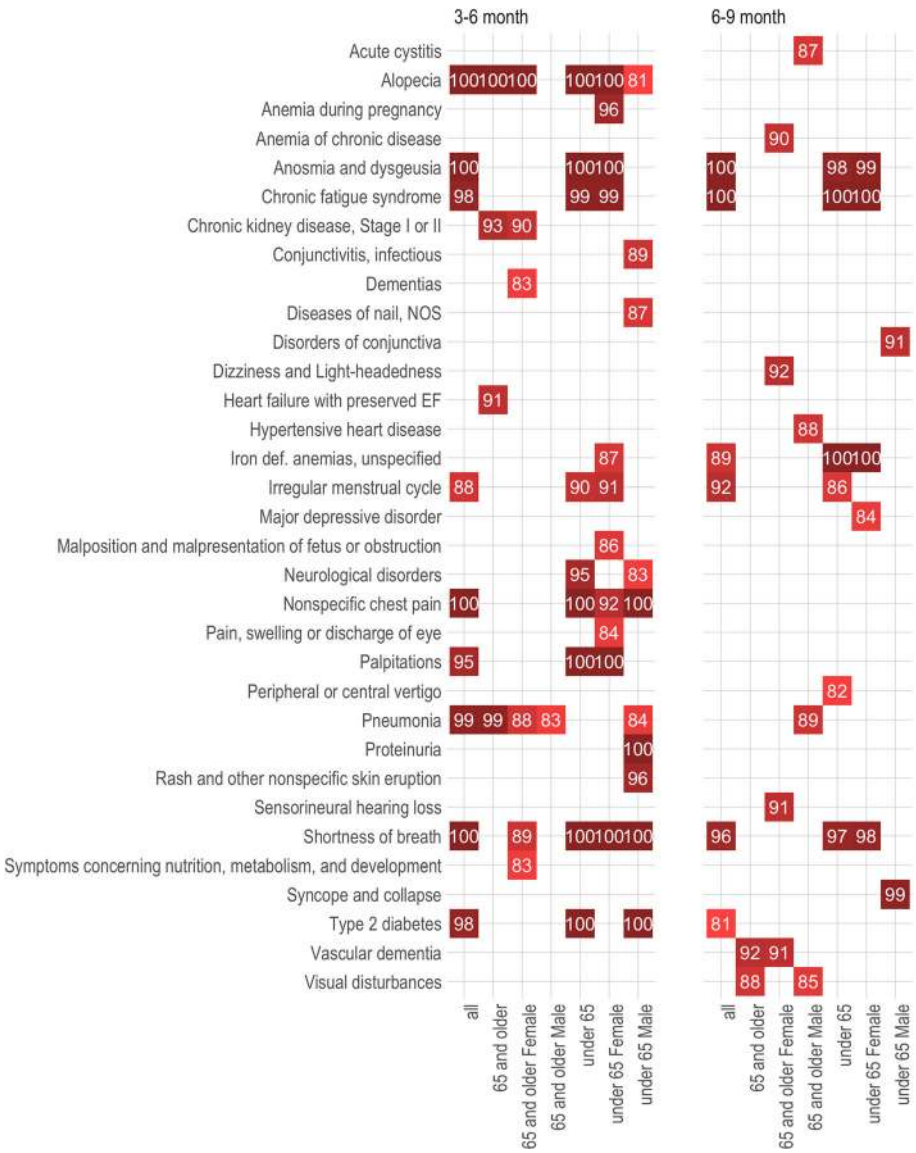

Fig. 2 Phenotypes that are positively associated with a past COVID-19-positive RT-PCR test. Identified post-COVID-19 phenotypes by age and gender and ordered by MLHO confidence scores (plotted in white font). One hundred means phenotype was identified in 100% of MLHO iterations. Phenotypes included have been associated with a positive past COVID-19 test with a confidence score higher than 80% in at least a sub-cohort

findings have important implications for younger patients. Despite having not been hospitalized during the acute phase, the symptoms of long COVID are found with high confidence in this younger cohort population. This gives another reason for young patients to opt for having the vaccination since the long-term effects of the disease are clearly not limited to older patients. While the precise biological causes of the sequelae are still unknown and under investigation, the enrichment of these diagnoses among younger cohorts may indicate that the robustness of the immune response in these patients is driving some of the post-COVID sequelae. However, these results should be understood and qualified in the context that, on average, younger patients who are often healthier than 65 and older have fewer interactions with healthcare systems (and thus fewer diagnosis records), which may lead to greater ease in detecting a signal in this younger cohort compared to an older cohort.

While the chart review's primary purpose was to determine if the clinical notes were in agreement with the ICD-9/10 labels, the reviewer also noted that physicians consistently attributed two of the phenotypes (alopecia, and anosmia and dysgeusia) to a previous history of COVID-19, whereas the physicians' notes did not

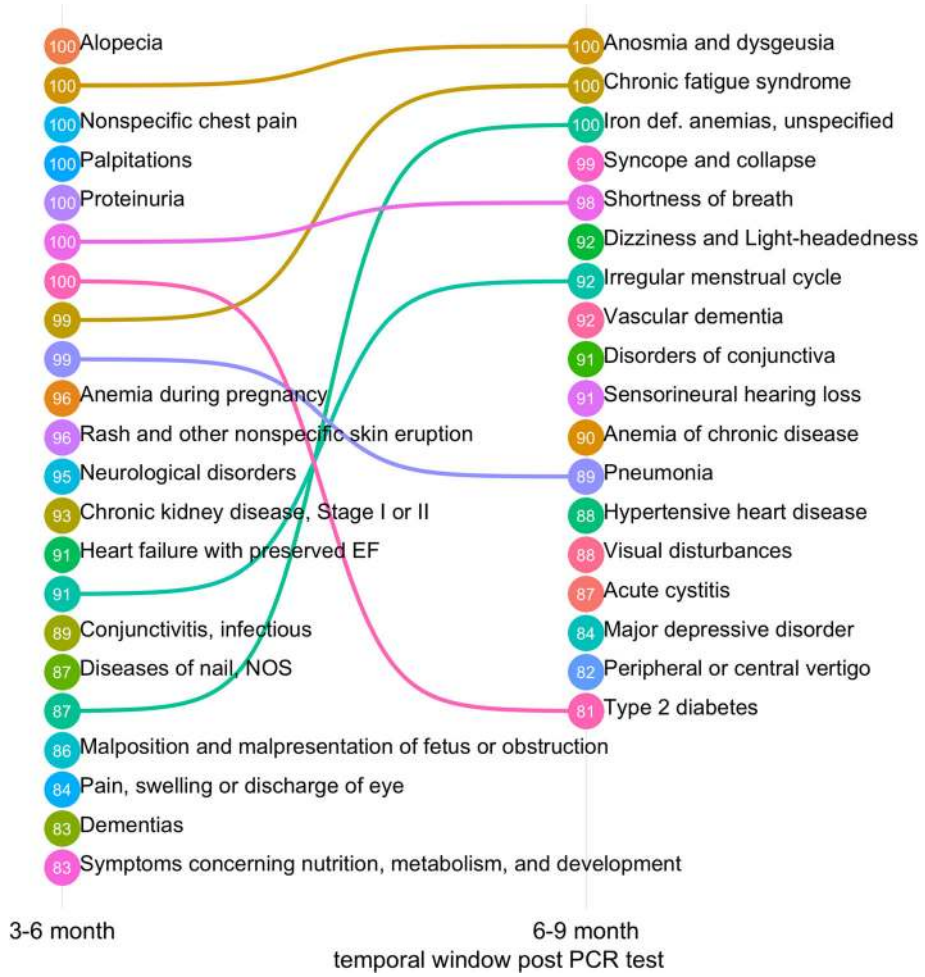

Fig. 3 Temporal presentation of PASC phenotypes 3–6 and 6–9 months post-COVID-19 RT-PCR test. PASC phenotypes are ranked by MLHO confidence score

specifically identify a connection between the phenotype and the previous infection for most of the other phenotypes, even those with high confidence like type 2 diabetes or non-specific chest pain. Our model indicates that even if these phenotypes are not explicitly identified or recognized by the clinician and patient at the individual level, many of these unrecognized phenotypes still have a high confidence score. While an ICD code on its own does not specify the time of onset, the chart review helped to confirm that the presented phenotypes were likely new since COVID-19. The majority of charts reviewed for each phenotype suggest that the symptoms or the diagnosis occurred after COVID-19. Our model identifies the relationships between COVID and a phenotype, where a healthcare provider and patient may otherwise miss that relationship.

Several neurological phenotypes (vascular dementia, dementia, and neurological disorders) were frequently diagnosed after COVID and appear to have an increased association with the infection. The neurological disorder phenotype includes several ICD codes, and in a random sampling of patients with this phenotype, the majority had the ICD code "R41.89—other symptoms and signs involving cognitive function and awareness." Collectively, these phenotypes suggest ongoing cognitive dysfunction. The earliest reports of acute COVID, such as Mao's retrospective analysis of 214 hospitalized patients in China, described neurological manifestations, including cerebrovascular complications, in nearly half of those with severe disease [37]. Since the acute phase, the sequelae for the description of "brain fog" after the diagnosis of COVID have been repeatedly described [38, 39]. Al-Aly specifically documents increased memory problems and strokes [40]. Our model suggests that these cognitive deficits are ongoing and in some cases may be so severe they are even lead to an initial formal diagnosis of dementia at higher rates among those with a history of COVID. While many of these patients may have

already shown some signs of memory loss, the formal diagnosis of dementia did not come until after COVID-19 suggesting that the viral illness may have contributed to a worsening of their condition and the formal declaration of this diagnosis.

Another important phenotype identified was type 2 diabetes. Several studies have pointed out possible pathophysiological relationships between COVID-19 and diabetes [41, 42]. And the increased incidence of a number of metabolic diseases has been found with those after a COVID-19 diagnosis [40]. Our study indicates that the metabolic disorder may be so significant as to lead to a formal diagnosis of diabetes mellitus.

The disease of the nail phenotype includes a variety of diagnoses including leukonychia, onycholysis, onychomadesis, Mees' lines, Muehrcke's lines, and Beau's lines all of which are markers of overall well-being and have been associated with infections and renal or hepatic dysfunction previously. Beau's lines have specifically been associated with COVID-19 infections [43, 44]. Our results suggest this association is widespread and likely a result of systemic infection including renal injury.

Proteinuria was also identified as having an association with COVID-19 among male patients less than 65. COVID-19 has previously been associated with acute kidney injury [45], and proteinuria is a known surrogate for kidney disease [46]. The identification of proteinuria as an association with COVID-19 in the young patient cohort suggests the insult of COVID-19 to the kidneys persists months after the infection has resolved.

The MLHO framework appears to be more powerful than univariate PheWAS. A small number of phenotypes that had a relatively high unadjusted statistical significance (a p-value between 0.01 and 0.001) would have been dropped in a linear univariate PheWAS after p-value correction for multiple hypotheses. Two examples of such phenotypes are palpitations and non-specific chest pain, both of which have previously been described as common symptoms of PASC [7, 35, 36].

MLHO's implementation in this study is similar to the standard univariate PheWAS [28, 29] as both offer computational solutions for high-throughput association mining from clinical data. However, a challenge in standard PheWAS is to find a sensible balance between adequately applying a correction to p-values in order to reduce false discovery due to multiple testing and minimizing false negatives [47]. Our approach expands the univariate p-value dependent criteria for identifying phenome-wide associations to a more comprehensive and multivariate entropy-based process. MLHO iteratively applies joint mutual information, performs sparsity screening, and uses gradient boosting to characterize the post-acute sequelae of COVID-19. The iterative process in MLHO provides means to an interpretable probabilistic confidence score for each phenotype associated with a past positive COVID-19 RT-PCR test.

Augmented with clinical expertise (i.e., chart reviews), MLHO's computational algorithms avoid a flood of false-positive discoveries while offering a more robust probabilistic approach than the standard PheWAS. We were able to evaluate over 1600 phenotypes and identify a small number of phenotypes (with confidence scores) that associate with a past COVID-19 infection. As a result, and along with the inclusion of COVID-negative patients, this study rules out some of the phenotype associations, which were previously identified through poorly controlled observational data, such as cutaneous eruptions outside of nail changes and alopecia.

We acknowledge that this study's findings may present limitations due to the use of only diagnosis codes, which can result in missing signs and symptoms that are in clinical notes and laboratory results. In addition, given the intensity of the pandemic and spread of misinformation, EHR data may represent confirmatory bias between providers and patients. Replicating this study in other institutions would help elucidate if the clinical phenotypes seen at MGB reflect true characteristics of PASC or local healthcare utilization patterns. Additionally, we only included diagnoses that were used for the first time at least 2 months after the COVID-positive PCR date. This may have led to some missed diagnoses that began within 2 months of the start of the acute phase; however, it helps ensure that the new diagnoses detected were not related to the acute phase. Future studies can consider modifying this time buffer; however, there will remain a trade-off between capturing all subsequent diagnoses and increasing the confidence that the diagnoses are not part of the acute phase of the illness. Finally, we have excluded hospitalized COVID-19 patients. On the one hand, it would be difficult to match hospitalized coronavirus patients during the COVID era with non-COVID hospitalized patients. On the other hand, the post-COVID syndrome can still be observed in patients who were never hospitalized [12, 48–52]. Regardless, future PASC studies should include hospitalized patients.

Conclusion

The COVID-19 pandemic in the USA raged nearly uncontrolled in 2020. While the exact number of people afflicted by the post-acute sequelae of SARS-CoV-2 infection is unknown, it represents a significant public health burden because of the large magnitude of the COVID-19 spread globally. We identified 33 phenotypes that were indicative of long COVID among non-hospitalized COVID-19 patients. Our understanding of COVID-19 and its chronic sequelae is evolving, and new risks are unknown. We do not know who might develop the post-COVID syndrome, how long the symptoms last,

and whether COVID-19 prompts the presentation of chronic diseases. Accurate identification of phenotypes will be important to guide future research and the healthcare system to focus its efforts and resources on adequately controlled age- and gender-specific sequelae of a COVID-19 infection. The ever-increasing adoption and magnitude of clinical data stored in EHR repositories over the past decade provide exceptional opportunities for instrumenting healthcare systems to study evolving pandemic byproducts. EHR data offer a unique opportunity to understand the post-acute effects that can follow SARS-CoV-2 infection.

Abbreviations

CI: Confidence interval; CS: Confidence score; EHRs: Electronic health records; ICD-10: 10th Revision of the International Statistical Classification of Diseases and Related Health Problems; ICD-9: 9th Revision of the International Statistical Classification of Diseases and Related Health Problems; MGB: Mass General Brigham; MSMR: Minimize sparsity, maximize relevance; OR: Odds ratio; PASC: Post-acute sequelae of COVID-19; PheWAS: Phenome-wide association studies; RT-PCR: Reverse transcription polymerase chain reaction

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-021-02115-0.

Additional file 1: Figure S1. Schema for counting diagnosis records in the cases and controls. **Figure S2.** Patient population selection. **Table S1.** Demographic characteristics of the study cohort. **Table S2.** Manual chart review of the 42 phenotypes identified by MHLO. **Table S3.** Multivariate ORs for PASC phenotypes [53–60].

Acknowledgements

We thank many colleagues in the Mass General Brigham Research Information Science & Computing team for curating MGB COVID-19 mart and providing information science and computing support. The members of the Consortium for Clinical Characterization of COVID-19 By EHR (4 CE) are as follows: James R Aaron, Giuseppe Agapito, Adem Albayrak, Mario Alessiani, Danilo F Amendola, Li L.L.J Anthony, Bruce J Aronow, Fatima Ashraf, Andrew Atz, Paul Avillach, James Balshi, Brett K Beaulieu-Jones, Douglas S Bell, Antonio Bellasi, Riccardo Bellazzi, Vincent Benoit, Michele Beraghi, José Luis Bernal Sobrino, Mélodie Bernaux, Romain Bey, Alvar Blanco Martínez, Martin Boeker, Clara-Lea Bonzel, John Booth, Silvano Bosari, Florence T Bourgeois, Robert L Bradford, Gabriel A Brat, Stéphane Bréant, Nicholas W Brown, William A Bryant, Mauro Bucalo, Anita Burgun, Tianxi Cai, Mario Cannataro, Aldo Carmona, Charlotte Caucheteux, Julien Champ, Jin Chen, Krista Chen, Luca Chiovato, Lorenzo Chiudinelli, Kelly Cho, James J Cimino, Tiago K Colicchio, Sylvie Cormont, Sébastien Cossin, Jean B Craig, Juan Luis Cruz Bermúdez, Jaime Cruz Rojo, Arianna Dagliati, Mohamad Daniar, Christel Daniel, Anahita Davoudi, Batsal Devkota, Julien Dubiel, Loic Esteve, Hossein Estiri, Shirley Fan, Robert W Follett, Paula S. A Gaiolla, Thomas Ganslandt, Noelia García Barrio, Lana X Garmire, Nils Gehlenborg, Alon Geva, Tobias Gradinger, Alexandre Gramfort, Romain Griffier, Nicolas Griffon, Olivier Grisel, Alba Gutiérrez-Sacristán, David A Hanauer, Christian Haverkamp, Bing He, Darren W Henderson, Martin Hilka, John H Holmes, Chuan Hong, Petar Horki, Kenneth M Huling, Meghan R Hutch, Richard W Issitt, Anne Sophie Jannot, Vianney Jouhet, Mark S Keller, Katie Kirchoff, Jeffrey G Klann, Isaac S Kohane, Ian D Krantz, Detlef Kraska, Ashok K Krishnamurthy, Sehi L'Yi, Trang T Le, Judith Leblanc, Andressa RR Leite, Guillaume Lemaitre, Leslie Lenert, Damien Leprovost, Molei Liu, Ne Hooi Will Loh, Sara Lozano-Zahonero, Yuan Luo, Kristine E Lynch, Sadiqa Mahmood, Sarah Maidlow, Alberto Malovini, Kenneth D Mandl, Chengsheng Mao, Anupama Maram, Patricia Martel, Aaron J Masino, Maria Mazzitelli, Arthur Mensch, Marianna Milano, Marcos F Minicucci, Bertrand Moal, Jason H Moore, Cinta Moraleda, Jeffrey S Morris, Michele Morris, Karyn L Moshal, Sajad Mousavi, Danielle L Mowery, Douglas A Murad, Shawn N Murphy, Thomas P Naughton, Antoine Neuraz, Kee Yuan Ngiam, James B Norman, Jihad Obeid, Marina P Okoshi, Karen L Olson, Gilbert S Omenn, Nina Orlova, Brian D Ostasiewski, Nathan P Palmer, Nicolas Paris, Lav P Patel, Miguel Pedrera Jimenez, Emily R Pfaff, Danielle Pillion, Hans U Prokosch, Robson A Prudente, Víctor Quirós González, Rachel B Ramoni, Maryna Raskin, Siegbert Rieg, Gustavo Roig Domínguez, Pablo Rojo, Carlos Sáez, Elisa Salamanca, Malarkodi J Samayamuthu, Arnaud Sandrin, Janaina CC Santos, Maria Savino, Emily R Schriver, Petra Schubert, Juergen Schuettler, Luigia Scudeller, Neil J Sebire, Pablo Serrano Balazote, Patricia Serre, Arnaud Serret-Larmande, Zahra Shakeri, Domenick Silvio, Piotr Sliz, Jiyeon Son, Charles Sonday, Andrew M South, Anastasia Spiridou, Amelia LM Tan, Bryce WQ Tan, Byorn WL Tan, Suzana E Tanni, Deanne M Taylor, Ana I Terriza Torres, Valentina Tibollo, Patric Tippmann, Carlo Torti, Enrico M Trecarichi, Yi-Ju Tseng, Andrew K Vallejos, Gael Varoquaux, Margaret E Vella, Guillaume Verdy, Jill-Jênn Vie, Shyam Visweswaran, Michele Vitacca, Kavishwar B Wagholikar, Lemuel R Waitman, Xuan Wang, Demian Wassermann, Griffin M Weber, Zongqi Xia, Nadir Yehya, William Yuan, Alberto Zambelli, Harrison G Zhang, Daniel Zoeller, and Chiara Zucco.

Authors' contributions

HE, ZHS, GAB, and SNM designed the study. HE acquired the data and performed the analyses. HE, ZHS, GAB, YRS, CJP, and SNM were involved in the interpretation of the results. HE and ZHS are the co-lead authors. HE, ZHS, GAB, YRS, CJP, and SNM contributed to the critical review of the manuscript. The authors read and approved the final draft.

Funding

This work was supported by the National Human Genome Research Institute grant 3U01HG008685-05S2 and the National Library of Medicine grant T15LM007092. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH nor Massachusetts General Hospital.

Availability of data and materials

The anonymized patient-level data used for this project cannot be shared for reasons of information governance. Data may be available to affiliated researchers given the MGB IRB approval. Computer code for MLHO is publicly available on GitHub: https://github.com/hestiri/mlho/

Declarations

Ethics approval and consent to participate

The use of clinical data in this study was approved by the MGB Institutional Review Board with a waiver of informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA 02114, USA. [2]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Department of Dermatology, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA. [6]Research Information Science and Computing, Mass General Brigham, Boston, MA, USA.

Received: 28 April 2021 Accepted: 1 September 2021
Published online: 27 September 2021

References

1. Carfì A, Bernabei R, Landi F. Others. Gemelli against COVID-19 post-acute care study group. Persistent symptoms in patients after acute COVID-19. JAMA. 2020;324:603–5.
2. Xiong Q, Xu M, Li J, Liu Y, Zhang J, Xu Y, et al. Clinical sequelae of COVID-19 survivors in Wuhan, China: a single-centre longitudinal study. Clin Microbiol Infect. 2021;27:89–95.

3. Goërtz YMJ, Van Herck M, Delbressine JM, Vaes AW, Meys R, Machado FVC, et al. Persistent symptoms 3 months after a SARS-CoV-2 infection: the post-COVID-19 syndrome? ERJ Open Res. 2020;6. https://doi.org/10.1183/2312 0541.00542-2020.
4. Meini S, Suardi LR, Busoni M, Roberts AT, Fortini A. Olfactory and gustatory dysfunctions in 100 patients hospitalized for COVID-19: sex differences and recovery time in real-life. Eur Arch Otorhinolaryngol. 2020;277:3519–23.
5. Halpin SJ, McIvor C, Whyatt G, Adams A, Harvey O, McLean L, et al. Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: a cross-sectional evaluation. J Med Virol. 2021;93:1013–22.
6. Bowles KH, McDonald M, Barrón Y, Kennedy E, O'Connor M, Mikkelsen M. Surviving COVID-19 After Hospital Discharge: Symptom, Functional, and Adverse Outcomes of Home Health Recipients. Ann Intern Med. 2020. https://doi.org/10.7326/M20-5206.
7. Nehme M, Braillard O, Alcoba G, Aebischer Perone S, Courvoisier D, Chappuis F, et al. COVID-19 Symptoms: Longitudinal eEvolution and Persistence in oOutpatient Settings. Ann Intern Med. 2020. https://doi.org/1 0.7326/M20-5926.
8. CDC. Long-Term Effects of COVID-19. 2020. https://www.cdc.gov/corona virus/2019-ncov/long-term-effects.html. Accessed 11 Mar 2021.
9. Hopkins C, Surda P, Whitehead E, Nirmal KB. Early recovery following new onset anosmia during the COVID-19 pandemic – an observational cohort study. J Otolaryngol Head Neck Surg. 2020;49. https://doi.org/10.1186/s404 63-020-00423-8.
10. Wong AW, Shah AS, Johnston JC, Carlsten C, Ryerson CJ. Patient-reported outcome measures after COVID-19: a prospective cohort study. Eur Respir J. 2020;56. https://doi.org/10.1183/13993003.03276-2020.
11. Carfì A, Bernabei R, Landi F. Gemelli Against COVID-19 Post-Acute Care Study Group. Persistent Symptoms in Patients After Acute COVID-19. JAMA. 2020;324:603–5.
12. Chopra V, Flanders SA, O'Malley M, Malani AN, Prescott HC. Sixty-day Outcomes Among Patients Hospitalized with COVID-19. Ann Intern Med. 2020. https://doi.org/10.7326/M20-5661.
13. Huang C, Huang L, Wang Y, Li X, Ren L, Gu X, et al. 6-Month consequences of COVID-19 in patients discharged from hospital: a cohort study. Lancet. 2021;397:220–32.
14. Garcia J, Gustavson AR. The science of self-report. APS Obs. 1997;10 https://www.psychologicalscience.org/observer/the-science-of-self-report.
15. Bakhoum MF, Ritter M, Garg AK, Chan AX, Bakhoum CY, Smith DM. Subclinical ocular inflammation in persons recovered from ambulatory COVID-19. medRxiv. 2020;2020.09.22.20128140.
16. Puntmann VO, Carerj ML, Wieters I, Fahim M, Arendt C, Hoffmann J, et al. Outcomes of Cardiovascular Magnetic Resonance Imaging in Patients Recently Recovered from Coronavirus Disease 2019 (COVID-19). JAMA Cardiol. 2020;5:1265–73.
17. Brito D, Meester S, Yanamala N, Patel HB, Balcik BJ, Casaclang-Verzosa G, et al. High Prevalence of Pericardial Involvement in College Student Athletes Recovering from COVID-19. JACC Cardiovasc Imaging. 2021;14:541–55.
18. Dani M, Dirksen A, Taraborrelli P, Torocastro M, Panagopoulos D, Sutton R, et al. Autonomic dysfunction in "long COVID": rationale, physiology and management strategies. Clin Med . 2021;21:e63–e67.
19. Gregorova M, Morse D, Brignoli T, Steventon J, Hamilton F, Albur M, et al. Post-acute COVID-19 associated with evidence of bystander T-cell activation and a recurring antibiotic-resistant bacterial pneumonia. Elife. 2020;9. https://doi.org/10.7554/eLife.63430.
20. Manckoundia P, Franon E. Is pPersistent Tthick cCopious Mmucus a lLong-Tterm Ssymptom of COVID-19? Eur J Case Rep Intern Med. 2020;7:002145.
21. Lu Y, Li X, Geng D, Mei N, Wu P-Y, Huang C-C, et al. Cerebral Mmicro-sStructural Cchanges in COVID-19 Ppatients - Aan MRI-based 3-month fFollow-up Sstudy. EClinicalMedicine. 2020;25:100484.
22. Raahimi MM, Kane A, Moore CE, Alareed AW. Late onset of Guillain-Barré syndrome following SARS-CoV-2 infection: part of "long COVID-19 syndrome"? BMJ Case Rep. 2021;14. https://doi.org/10.1136/bcr-2020-240178.
23. Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. J Am Med Inform Assoc. 2020. https://doi.org/10.1093/jamia/ocaa288.
24. Estiri H, Vasey S, Murphy SN. Transitive sequential pattern mining for discrete clinical data. International Conference on Artificial. 2020. https://link.springer.com/chapter/10.1007/978-3-030-59137-3_37.
25. Estiri H, Strasser ZH, Klann JG, McCoy TH Jr, Wagholikar KB, Vasey S, Castro VM, Murphy ME, Murphy SN. Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations. Patterns (N Y). 2020;1(4):100051.
26. Estiri H, Strasser ZH, Murphy SN. Individualized prediction of COVID-19 adverse outcomes with MLHO. Sci Rep. 2021;11:5322.
27. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN. Predicting COVID-19 mortality with electronic medical records. NPJ Digit Med. 2021;4:15.
28. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26:1205–10. https://doi.org/10.1093/bioinformatics/btq126.
29. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. Annu Rev Genomics Hum Genet. 2016; 17:353–73.
30. POSTDICTION. https://www.lexico.com/en/definition/postdiction. Accessed 2 Jul 2021.
31. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN. Predicting COVID-19 Mortality with Electronic Medical Records. npj Digital Medicine. https://doi.org/10.1038/s41746-021-00383-x.
32. Estiri H, Vasey S, Murphy SN. Transitive Sequential Pattern Mining for Discrete Clinical Data. In: Artificial intelligence in medicine. Springer International Publishing; 2020. p. 414–424.
33. Estiri H, Vasey S, Murphy SN. Generative transfer learning for measuring plausibility of EHR diagnosis records. J Am Med Inform Assoc. 2020. https://doi.org/10.1093/jamia/ocaa215.
34. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4: 1244.
35. Nalbandian A, Sehgal K, Gupta A, Madhavan MV, McGroder C, Stevens JS, et al. Post-acute COVID-19 syndrome. Nat Med. 2021. https://doi.org/10.103 8/s41591-021-01283-z.
36. Petersen MS, Kristiansen MF, Hanusson KD, Danielsen ME, Á Steig B, Gaini S, et al. Long COVID in the Faroe Islands - a longitudinal study among non-hospitalized patients. Clin Infect Dis. 2020. doi:https://doi.org/10.1093/cid/ciaa1792.
37. Mao L, Jin H, Wang M, Hu Y, Chen S, He Q, et al. Neurologic Manifestations of Hospitalized Patients with Coronavirus Disease 2019 in Wuhan. China. JAMA Neurol. 2020;77:683–90.
38. Graham EL, Clark JR, Orban ZS, Lim PH, Szymanski AL, Taylor C, et al. Persistent neurologic symptoms and cognitive dysfunction in non-hospitalized COVID-19 "long haulers." Ann Clin Transl Neurol. 2021;8:1073–1085.
39. Bell ML, Catalfamo CJ, Farland LV, Ernst KC, Jacobs ET, Klimentidis YC, et al. Post-acute sequelae of COVID-19 in a non-hospitalized cohort: results from the Arizona CoVHORT. bioRxiv. 2021. doi:https://doi.org/10.1101/2021.03.2 9.21254588.
40. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature. 2021. https://doi.org/10.1038/s41586-021-03 553-9.
41. Maddaloni E, Buzzetti R. Covid-19 and diabetes mellitus: unveiling the interaction of two pandemics. Diabetes Metab Res Rev. 2020;:e33213321.
42. Hayden MR. An Immediate and Long-Term Complication of COVID-19 May be Type 2 Diabetes Mellitus: the Central Role of β-Cell Dysfunction, Apoptosis and Exploration of Possible Mechanisms. Cells. 2020;9. https://doi.org/10.3390/cells9112475.
43. Alobaida S, Lam JM. Beau lines associated with COVID-19. CMAJ. 2020;192: E1040.
44. Ide S, Morioka S, Inada M, Ohmagari N. Beau's Lines and Leukonychia in a COVID-19 Patient. Intern Med. 2020;59:3259.
45. Rudnick MR, Hilburg R. Acute Kidney Injury in COVID-19: Another Challenge for Nephrology. Am J Nephrol. 2020;51:761–3.
46. Cravedi P, Remuzzi G. Pathophysiology of proteinuria and its value as an outcome measure in chronic kidney disease. Br J Clin Pharmacol. 2013;76:516–23.
47. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100:9440–5.
48. Wijeratne T, Crewther S. Post-COVID 19 Neurological Syndrome (PCNS); a novel syndrome with challenges for the global neurology community. J Neurol Sci. 2020;419:117179.
49. Garg P, Arora U, Kumar A, Wig N. The "post-COVID" syndrome: how deep is the damage? J Med Virol. 2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.26465.

50. Marshall M. The lasting misery of coronavirus long-haulers. Nature. 2020;585:339–41.
51. Rubin R. As their Nnumbers Grow, COVID-19 "Long Haulers" Stump Experts. JAMA. 2020;324:1381–3.
52. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of Llong-COVID: analysis of COVID cases and their symptoms collected by the COVID Symptoms Study App. bioRxiv. 2020. doi: https://doi.org/10.1101/2020.10.19.20214494.
53. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One. 2017;12:e0175508.
54. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform. 2019;7:e14325.
55. Bennasar M, Hicks Y, Setchi R. Feature Selection using Joint Mutual Information Maximisation. Expert Syst Appl. 2015;42:8520–32.
56. Yang H, Moody J. Feature selection based on joint mutual information. In: Proceedings of linternational ICSC symposium on advances in intelligent data analysis. Citeseer; 1999. p. 22–25.
57. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based Boosting in R. Comput Stat. 2012. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.2755&rep=rep1&type=pdf.
58. Hofner B, Hothorn T, Kneib T, Schmid M. A Framework for Unbiased Model Selection Based on Boosting. J Comput Graph Stat. 2011;20:956–71.
59. Bühlmann P, Hothorn T. Boosting Algorithms: Regularization, Prediction, and Model Fitting. Stat Sci. 2007;22:477–505.
60. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: a hands-on tutorial using the R package mboost. Comput Stat. 2014;29:3–35.

Publisher's Note