Written Testimony for the U.S. Senate Committee on Health, Education, Labor and Pensions hearing on
Addressing Long COVID:
Advancing Research and Improving Patient Care

Charisse Madlock-Brown, PhD
Associate Professor of Health Informatics
Acute and Critical Care Division
College of Nursing
University of Iowa
Iowa City, IA

My Background

Chairman Sanders, Ranking Member Cassidy, and Members of the Committee, thank you for the opportunity to participant in this hearing. I am a faculty member in Health Informatics at the University of Iowa College of Nursing. I was previously an associate professor at the University of Tennessee Health Science Center. I received my Master's in Library and Information Science and Ph.D. in Health Informatics from the University of Iowa. I have a broad background in health informatics, with a current focus on social determinants of health, COVID-19, health disparities, obesity trends, and multimorbidity. I use machine learning and bio-statistics to analyze large electronic health record (EHR) data warehouses. Since 2021, I have been the co-lead for the National Cohort Cohort Collaborative's (N3C) Social Determinant of Health (SDoH) domain team directing research on SDoH and COVID-19 outcomes as well as data engineering efforts to improve harmonization of individual level SDoH data into Data warehouses. Additionally, I am the Iowa site PI for the Center for Linkage and Acquisition of Data (CLAD) for the All of Us program.[1] The "All of Us" Research Program is a significant initiative aiming to recruit at least one million participants from across the diverse spectrum of the United States. CLAD will further the program's efforts through integrating claims, mortality, and environmental data into existing EHR and survey data.

The N3C project,[1] developed in partnership with the National Center for Advancing Translational Sciences (NCATS), focuses on identifying cases of COVID-19 in a network of HER systems, including those that are laboratory-confirmed, suspected, or considered possible. To facilitate accurate comparisons and analyses, these COVID-19 cases are demographically matched with control subjects who have tested negative or equivocal for COVID-19. The matching criteria include age group, sex, race, and ethnicity. The ratio of COVID-19 cases to control subjects is maintained at 1:2, ensuring a robust and representative comparison group for the study. N3C continuously grows its network and currently comprises 83 distinct healthcare sites representing all regions in the U.S. and 21.7 million patients. N3C is part of the EHR arm of the Researching COVID to Enhance Recovery (RECOVER) NIH initiative.[2] As a funded N3C RECOVER researcher, I work with a diverse team of clinicians, patient representatives, epidemiologists, and informatics professionals to investigate patterns in EHR data related to long COVID.

These statements are made on my behalf and do not represent N3C or RECOVER.

Benefits of leveraging EHR database systems for long COVID research

Large-scale Electronic Health Record (EHR) warehouse networks have played a pivotal role in rapidly developing insights into long COVID due to several key factors. Firstly, EHRs contain a breadth of coded patient health information, including histories of diagnoses, medication prescriptions, lab results, vital measurements and outcomes.[3] Additionally, EHRs cover large and diverse patient populations across various demographics, geographic locations, and healthcare settings. This diversity is crucial for studying long COVID in different groups, enhancing the generalizability of findings.[1] Another significant advantage of EHRs is the

provision of longitudinal data, allowing researchers to track patients' health over extended periods.[4] This aspect is particularly important for long COVID, characterized by prolonged and evolving symptoms. Inquiries into observational EHR data warehouses can be launched immediately, provided the data is available. In contrast, clinical trials take a great deal of time to launch, enroll, and complete prior to the data becoming available for analysis. This real-world data aids in understanding patterns of clinical management surrounding long COVID.[6] EHRs also contain critical data on comorbidities,[6] shedding light on risk factors and the impact of long COVID on individuals with specific health profiles. EHRs also enable the identification of subgroups within long COVID patients, presenting distinct clinical features, which is key for understanding a disease that may have more than one etiology and presentation.[8] The ability to share EHR databases across institutions facilitates collaborative research and larger-scale studies[1], essential for understanding complex conditions like long COVID.

The continuous updating of EHRs ensures rapid availability of recent data - crucial for an emerging condition like long COVID where knowledge is rapidly evolving. Regarding pharmacovigilance, EHR data can be instrumental in monitoring adverse drug reactions,[8] which may be particularly important as new treatments for long COVID emerge. Lastly, using existing EHR data is often more cost-effective than new data collection,[9] especially relevant for a widespread condition like long COVID. This cost-effectiveness makes EHRs an invaluable resource in the ongoing research and understanding of long COVID.

Challenges Associated with Leveraging EHR databases for research.

Numerous challenges are associated with using EHR data for research given limitations such as the inherently incomplete nature of patient charts.[10] Events like visits at external clinics or home COVID tests likely won't be captured. This type of incompleteness can be partially mitigated through patient privacy preserving linkage between different healthcare systems as well as billing data sources,[11] such as CMS. Variations in data capture and data mappings between electronic systems also impact data incompleteness. For example, long COVID diagnosis coding varied over time and by site,[12] as did the mapping of this code to research data warehouses. Implementing data harmonization pipelines and quality assessments within EHR analytic platforms can enhance data interoperability and inform data selection and cleaning prior to analysis.[3]

Observational health researchers much also keep in mind that patients who visit large healthcare systems are not representative of the entire population as healthy patients or patients with less access to care will be less present in the sample.[13] This is partially mitigated by including a diverse range of healthcare sites and linking CMS data from all patient encounters so that researchers can ensure a more representative sample that includes patients with milder illness and fewer healthcare interactions. Patient utilization differences by sub-groups due to varying levels of healthcare access among different demographic groups can be partially overcome by ensuring representation from diverse communities and analyzing utilization patterns to identify and correct for bias.

Benefits of large EHR networking efforts.

Large EHR networking efforts have helped researcher make progress on all mitigating factors related to the limitations of observational data.[14] The lack of standardization in EHR systems,[15] a significant barrier in aggregating and comparing data across sources, has been tackled by large networks implementing harmonization pipelines, and facilitating data standardization at the national level. A large EHR network can establish data quality benchmarks[14] by showing institutions where they might be lacking certain data richness when compared to other members of the network. This can support investigation and improvement of local data collection and mapping practices.

By addressing these various challenges through the strategic use of large EHR warehouse networks, research into long COVID and similar complex health conditions (e.g., Myalgic encephalomyelitis/chronic fatigue syndrome and Postural tachycardia syndrome), can become more accurate, inclusive, and representative, leading to more effective and personalized healthcare solutions.

Methods for Identifying long COVID patients

Identifying patients with long COVID in Electronic Health Record (EHR) systems presents a significant challenge in current medical research as the condition is underdiagnosed.

Machine Learning Models: Pfaff et al. developed a machine learning model to identify probable long COVID patients using a training dataset from 597 patients at long-COVID clinics using N3C.[16] The XGBoost model was tested in a cohort of 97,995 COVID-19 patients into three categories: all COVID-19 patients, hospitalized, and non-hospitalized patients. Key features for identifying long COVID included healthcare utilization, patient age, dyspnea, and other diagnosis and medication information. The models showed high accuracy in pinpointing potential long-COVID patients.

Prospective Observational Cohort Study: Researchers conducted a nationwide, prospective observational cohort study, involving adults who completed a symptom survey six months after acute COVID-19 symptom onset or testing positive.[17] The study identified PASC based on 44 self-reported symptoms, emphasizing symptoms like post exertional malaise, fatigue, and brain fog. This method provided a symptom-based definition of PASC.

Analysis of U09.9 ICD-10-CM Code Usage: Researchers used theICD-10-CM code U09.9, "Post COVID-19 condition, unspecified," in the N3C system to identify patients with long COVID.[12] It focused on 33,782 patients with a U09.9 diagnosis, examining demographics, social determinants of health, co-occurring diagnoses, medications, and procedures within 60 days of diagnosis, particularly noting age group differences. The study found that long COVID diagnoses (U09.9) were predominantly among female, White, non-Hispanic individuals from low poverty and unemployment areas, highlighting disparities in long COVID diagnosis and suggesting the need for further research and action.

Research using EHR data warehouses to identify symptoms and risk factors associated with long COVID

Research utilizing EHR data warehouses has significantly contributed to understanding the symptoms and risk factors of COVID-19 and its prolonged effects, often referred to as Long COVID or post-acute sequelae of SARS-CoV-2 infection (PASC). Sudre CH et al.'s review identified symptoms of long COVID common across multiple studies, including fatigue, exertional dyspnea, musculoskeletal pain, and "brain fog," with additional mental health issues like anxiety, depression, and PTSD.[18] This study emphasizes the broad range of symptoms that can persist or emerge well after the acute phase of the infection. A comprehensive study across 31 health systems in the United States, part of N3C, involved 8,325 individuals with PASC.[20] This study examined a range of factors, including demographics, comorbidities, and acute characteristics of COVID-19, to identify risk factors for PASC. Key findings included the higher prevalence of diagnosis of PASC in individuals over 50 years old, females, non-Hispanic Whites, and those with specific comorbidities like depression, chronic lung disease, and obesity. Zang et al used electronic health records from two large networks, INSIGHT and OneFlorida+, covering over 27 million patients in NYC and Florida, to investigate post-acute sequelae of SARS-CoV-2 infection (PASC, or long COVID).[20] Employing a high-throughput screening approach, the study identified a range of diagnoses and medications with higher incidence in patients 30-180 days post-COVID infection compared to non-infected individuals. Notable findings include a greater number of PASC cases in NYC than Florida, with conditions like dementia, hair loss, and pulmonary issues common in both cohorts. The study highlights the variability in long COVID risks across different populations. These studies, leveraging the extensive data available in EHR warehouses, underscore the diverse and complex nature of COVID-19 and its long-term effects. They highlight the critical role of comprehensive data analysis in understanding this evolving health issue, informing potential early intervention strategies, and guiding future clinical and epidemiological research for effective management and treatment of PASC.

Characterizing and Sub-Phenotyping long COVID using EHR data warehouses

Reese et al. stratified patients with PASC (or long COVID) through computational modeling of phenotype data from EHRs.[21] It assesses phenotypic similarities between patients using semantic similarity, identifying six distinct PASC patient clusters. Specific abnormalities, such as pulmonary, neuropsychiatric, and cardiovascular issues, characterize these clusters. The approach was validated across different hospital systems. Pfaff et al analyzed data from 33,782 patients diagnosed with U09.9 and identified four major categories of diagnoses that commonly co-occur with the U09.9 code.[12] These categories include cardiopulmonary, neurological, gastrointestinal, and comorbid conditions. This method provides insights into the prevalent health issues associated with long COVID and how they are documented in healthcare records. It also has evolved into new sub-phenotyping approaches used in N3C RECOVER queries. These papers show that, despite the varied nature of the symptomology of long COVID, it is possible to characterize symptoms in broad categories, which will help future researchers develop patient profiles for long COVID.

Long COVID and other Post-viral Syndromes

This section explores the intersection between Post-Acute Sequelae of SARS-CoV-2 infection (PASC or long COVID) and a spectrum of related chronic conditions, notably myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). The emergence of long COVID has cast a spotlight on these conditions, underscoring the urgent need for more research into their shared characteristics and underlying mechanisms.

Redox Imbalance in COVID-19 and ME/CFS: This review highlights the similarities between PASC and myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), particularly in terms of redox imbalance, systemic and neuroinflammation, impaired ATP generation, and hypometabolism.[22] These biological abnormalities provide evidence of a shared biological basis between long COVID-19 and ME/CFS, indicating potential pathways for new therapeutic approaches.

Long COVID vs. Long Flu in the Elderly: An observational cohort study compared long COVID with residual symptoms in elderly influenza patients (termed "long Flu").[23] The study found that long COVID patients exhibited a higher incidence of symptoms like dyspnea, fatigue, palpitations, loss of taste/smell, and neurocognitive symptoms compared to those with long Flu, suggesting differences in the symptomatology and severity between post-viral syndromes in elderly patients.

Long COVID as a Form of infection-associated chronic illness/ME/CFS: This paper posits that long COVID is essentially the same condition as ME/CFS, also an infection-associated chronic illness.[24] It argues that acute COVID-19 triggers ME/CFS, similar to other infectious agents. The literature review by Komaroff et al. notes considerable similarities between ME/CFS and Long COVID, particularly in symptomatology and biological abnormalities.[25] It emphasizes that ME/CFS often follows an infectious-like illness and is characterized by post-exertional malaise, similar to Long COVID. In fact, several studies estimate that half of long COVID patients fit the criteria for ME/CFS.[26]

Collectively, these papers suggest significant overlaps in the symptomatology and underlying biological abnormalities between PASC and other infection-associated chronic conditions, particularly ME/CFS, indicating a potential common pathophysiological basis.

Future directions

A moonshot initiative is urgently needed in the wake of the persistent and global challenge of long COVID.[27] The current landscape of research into this condition is alarmingly disjointed. Experts in diverse medical specialties, including pulmonology, neurology, and cardiology, remain siloed, rarely crossing paths to share insights. The scarcity of clinical trials aimed at the underlying causes of long COVID and the need for a robust infrastructure for swift trial implementation pose a significant barrier to progress.[28] This situation is further exacerbated by the inadequate long-term funding, causing hesitation among scientists and companies in pursuing potential treatments.[27]

This call to action urges the US government to spearhead this moonshot by committing an annual investment of at least $1 billion over the next ten years. Such a bold move could galvanize global efforts, encouraging governments worldwide to rise to this health challenge that affects every continent. The effectiveness of substantial funding is evident in cancer research as seen through initiatives like the Cancer Moonshot.[29]

Primary activities for a long COVID moonshot must be clinical trials so that effective treatments can be tested and developed. There is a particular need for studies on experimental medicines, as currently, there are only 12 such studies for long COVID listed in ClinicalTrials.gov[30]. In tandem with clinical trials, there are several ways observational research systems like N3C can both enhance our understanding of long COVID and bolster findings from trials:

1. **Identification of Symptom Sub-Phenotypes**: Research should aim to identify common symptom sub-phenotypes across various post-viral syndromes. This will enhance our understanding of long COVID and similar conditions, leading to more effective treatments and management strategies.
2. **Improving Diagnosis and Treatment Accessibility**: Addressing biases in long COVID diagnosis, as highlighted by research from the N3C RECOVER team, is critical. Efforts must focus on making diagnoses and treatments more accessible to all demographics, particularly those currently underdiagnosed. Furthermore, enhancements of SDoH data within these systems[31] can be further leveraged to understand biases in available data for patients with limited access to care and identification of health disparities in long COVID diagnoses and treatment across sociodemographic areas.
3. **Clinical Guidelines:** Establishing clinical guidelines to support the accurate diagnosis of long COVID is imperative[32]. These guidelines will ensure consistency and reliability in diagnosing this condition across various healthcare settings.
4. **Coordinated Clinical Programs for Underserved Communities**: Special attention must be given to coordinating clinical programs that specifically target underserved communities. This includes people in rural areas, low-income groups, and racial minorities who have limited access to healthcare.
5. **Structuring of Clinical Note and Flowsheet Data**: Ensuring that more data, such as SDoH data and complete symptom lists, become structured and easily accessible is vital for comprehensive research and analysis.
6. **Linking Prospective Data with Clinical Trials and Other Datasets**: More prospective data should be linked to clinical repositories using Privacy-Preserving Record Linkage (PPRL) to clinical trials data. Additionally, integrating this data with claims and mortality data will provide a more complete picture of various treatments' long-term impacts and effectiveness.

By focusing on these areas, the moonshot initiative can significantly advance our understanding of long COVID, leading to more effective treatments and better outcomes for those affected.

References

1. All of Us Research Program Establishes New Center for Linkage and Acquisition of Data. *All of Us Research Program | NIH* https://allofus.nih.gov/news-events/announcements/all-us-research-program-establishes-new-center-linkage-and-acquisition-data (2023).
2. Haendel, M. A. *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J. Am. Med. Inform. Assoc. JAMIA* **28**, 427–443 (2021).
3. National Institutes of Health. RECOVER: Researching COVID to Enhance Recovery. *RECOVER: Researching COVID to Enhance Recovery* https://recovercovid.org.
4. Bradwell, K. R. *et al.* Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (EHR) dataset. *J. Am. Med. Inform. Assoc.* **29**, 1172–1182 (2022).
5. Violán, C. *et al.* Five-year trajectories of multimorbidity patterns in an elderly Mediterranean population using Hidden Markov Models. *Sci. Rep.* **10**, 16879 (2020).
6. Jones, R. *et al.* Risk Predictors and Symptom Features of Long COVID Within a Broad Primary Care Patient Population Including Both Tested and Untested Patients. *Pragmatic Obs. Res.* **12**, 93–104 (2021).
7. Alshakhs, M., Jackson, B., Ikponmwosa, D., Reynolds, R. & Madlock-Brown, C. Multimorbidity patterns across race/ethnicity as stratified by age and obesity. *Sci. Rep.* **12**, 9716 (2022).
8. O'Neil, S. T. *et al.* Finding Long-COVID: Temporal Topic Modeling of Electronic Health Records from the N3C and RECOVER Programs. 2023.09.11.23295259 Preprint at https://doi.org/10.1101/2023.09.11.23295259 (2023).

9. Muzaffar, A. F., Abdul-Massih, S., Stevenson, J. M. & Alvarez-Arango, S. Use of the Electronic Health Record for Monitoring Adverse Drug Reactions. *Curr. Allergy Asthma Rep.* **23**, 417–426 (2023).
10. Mc Cord, K. A. *et al.* Current use and costs of electronic health records for clinical trial research: a descriptive study. *CMAJ Open* **7**, E23–E32 (2019).
11. Kim, E. *et al.* The Evolving Use of Electronic Health Records (EHR) for Research. *Semin. Radiat. Oncol.* **29**, 354–361 (2019).
12. Kiernan, D. *et al.* Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet®, the National Patient-Centered Clinical Research Network. *BMC Res. Notes* **15**, 337 (2022).
13. Pfaff, E. R. *et al.* Coding long COVID: characterizing a new disease through an ICD-10 lens. *BMC Med.* **21**, 58 (2023).
14. National Academies of Sciences, E., Division, H. and M., Services, B. on H. C. & Disabilities, C. on H. C. U. and A. with. *Factors That Affect Health-Care Utilization*. *Health-Care Utilization as a Proxy in Disability Determination* (National Academies Press (US), 2018).
15. Sidky, H. *et al.* Data quality considerations for evaluating COVID-19 treatments using real world data: learnings from the National COVID Cohort Collaborative (N3C). *BMC Med. Res. Methodol.* **23**, 46 (2023).
16. Weiskopf, N. G., Hripcsak, G., Swaminathan, S. & Weng, C. Defining and measuring completeness of electronic health records for secondary use. *J. Biomed. Inform.* **46**, 830–836 (2013).

17. Pfaff, E. R. *et al. Who has long-COVID? A big data approach*. 2021.10.18.21265168 https://www.medrxiv.org/content/10.1101/2021.10.18.21265168v1 (2021) doi:10.1101/2021.10.18.21265168.
18. Thaweethai, T. *et al.* Development of a Definition of Postacute Sequelae of SARS-CoV-2 Infection. *JAMA* **329**, 1934–1946 (2023).
19. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nat. Med.* (2021) doi:10.1038/s41591-021-01292-y.
20. Hill, E. *et al.* Risk Factors Associated with Post-Acute Sequelae of SARS-CoV-2 in an EHR Cohort: A National COVID Cohort Collaborative (N3C) Analysis as part of the NIH RECOVER program. 2022.08.15.22278603 Preprint at https://doi.org/10.1101/2022.08.15.22278603 (2022).
21. Zang, C. *et al.* Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative. *Nat. Commun.* **14**, 1948 (2023).
22. Reese, J. T. *et al.* Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. *EBioMedicine* **87**, 104413 (2023).
23. Paul, B. D., Lemle, M. D., Komaroff, A. L. & Snyder, S. H. Redox imbalance links COVID-19 and myalgic encephalomyelitis/chronic fatigue syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2024358118 (2021).
24. Fung, K. W., Baye, F., Baik, S. H., Zheng, Z. & McDonald, C. J. Prevalence and characteristics of long COVID in elderly patients: An observational cohort study of over 2 million adults in the US. *PLOS Med.* **20**, e1004194 (2023).

25. Williams, S. P., Michelle A. Long Covid is a new name for an old syndrome. *STAT* https://www.statnews.com/2023/09/14/long-covid-me-cfs-myalgic-encephalomyelitis-chronic-fatigue/ (2023).
26. Komaroff, A. L. & Lipkin, W. I. ME/CFS and Long COVID share similar symptoms and biological abnormalities: road map to the literature. *Front. Med.* **10**, 1187163 (2023).
27. Grach, S. L., Seltzer, J., Chon, T. Y. & Ganesh, R. Diagnosis and Management of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. *Mayo Clin. Proc.* **98**, 1544–1551 (2023).
28. McCorkell, L. & Peluso, M. J. Long COVID research risks losing momentum – we need a moonshot. *Nature* **622**, 457–460 (2023).
29. 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).
30. About the Cancer Moonshot^{SM} - NCI. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/about (2022).
31. Diseases, T. L. I. Where are the long COVID trials? *Lancet Infect. Dis.* **23**, 879 (2023).
32. Phuong, J. *et al.* Advancing Interoperability of Patient-level Social Determinants of Health Data to Support COVID-19 Research. *AMIA Summits Transl. Sci. Proc.* **2022**, 396–405 (2022).
33. Fernández-de-las-Peñas, C., Palacios-Ceña, D., Gómez-Mayordomo, V., Cuadrado, M. L. & Florencio, L. L. Defining Post-COVID Symptoms (Post-Acute COVID, Long COVID, Persistent Post-COVID): An Integrative Classification. *Int. J. Environ. Res. Public. Health* **18**, 2621 (2021).