Operationalizing the Undefined: How the Informatics Community Navigated Long COVID in N3C

Vibhor Gupta[1], Safoora Masoumi[1], Yue Wang PhD[1]
[1]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Introduction

Long COVID, or post-acute sequelae of SARS-CoV-2 infection (PASC), is a set of conditions that develop and persist after the initial infection of the COVID-19 virus. However, due to the complicated nature of long-term effects of COVID-19, it took the medical community four years (2020-2024) to arrive at an official definition of "Long COVID." In June 2024, the U.S. National Academies of Sciences, Engineering and Medicine (NASEM) published the report "*A Long Covid Definition*",[1] which was later adopted by the Centers for Disease Control and Prevention. The report defined Long COVID (LC) as "an infection-associated chronic condition (IACC) that occurs after SARS-CoV-2 infection and is present for at least 3 months as a continuous, relapsing and remitting, or progressive disease state that affects one or more organ systems." The report also discussed the process for NASEM to reach the definition and various characteristics of Long COVID.

The multi-year absence of an official Long COVID definition resulted in both practical and intellectual challenges. Without a clear definition to guide clinical practice, Long COVID patients may not be properly diagnosed or timely treated. The ambiguity also led to inconsistent coding and documentation of Long COVID in electronic health records, which made it challenging for data-driven methods to monitor, model, and understand Long COVID.

Encouragingly, the absence of a clear Long COVID definition did not stop the medical informatics community from using data to empirically study long COVID. Various criteria were proposed to operationalize "Long COVID" as a computable medical concept based on electronic health records. These empirical works gradually unveiled various symptoms and diseases associated with Long COVID, which contributed to the definitional characterization of this complex medical condition.

In this work, we review the provisional Long COVID definitions explored by the medical informatics community over multiple years, with a focus on empirical works using electronic health records in the National COVID Cohort Collaborative (N3C). In doing so, we aim to answer two research questions (RQs):

RQ1 (Descriptive): How did the medical informatics research community operationalize "Long COVID" in empirical data analysis? The evolution of long COVID definitions is an interesting phenomenon. It gives insights into how the medical informatics community tackled ambiguity and identified useful patterns from empirical data.

RQ2 (Prescriptive): What are some key lessons one can draw in retrospection which may accelerate the convergence process? It took many years for the global community to converge at a clear definition for this complex medical condition. It would be ideal if the convergence could happen sooner to improve care and minimize confusion.

Methods

A simple search of "Long Covid" in PubMed would yield more than 44,000 results, many of which are not empirical studies of Long COVID. We decided to focus on medical publications that operationalized "Long COVID" as a computable medical concept in one of the largest COVID-related research data repository, the National COVID Cohort Collaborative (N3C). We searched PubMed with the following query:

```
(("national covid cohort collaborative"[Title/Abstract]) OR (n3c[Title/Abstract])) AND (("long-covid"[Title/Abstract]) OR ("long covid"[Title/Abstract]) OR (long-haul COVID[Title/Abstract]) OR (Post-Acute COVID-19 syndrome[MeSH Terms]))
```

The query yielded 54 results as of January 2026. All results were screened manually to remove duplicates and papers without full text. A set of in-house inclusion/exclusion criteria were developed to focus on the studies that used N3C data and either defined or used Long COVID status for their research purposes. We excluded any commentary works or literature reviews on Long COVID. A total of 32 articles remained and were used for further analysis.

For each of the 32 articles, the computational criteria or procedure to define the binary status of whether the patient had Long COVID was extracted manually from the full text (usually in the "Methods" section). For example (next page):

A long COVID clinical diagnosis was defined by the addition of the long COVID International Classification of Diseases Tenth Revision (ICD-10) diagnostic code (U09.9) or a B94.8 ICD-10 diagnosis, which was used as a proxy for long COVID before the ICD-10 diagnostic code was created in October 2021.

Such a definition was then analyzed and parsed as a logical combination of vocabulary terms. For example, the above definition was parsed as {"ICD-10 code U09.9" OR "ICD-10 code B94.8"}, which contains two terms combined by a logical OR. The extracted data and analysis code is available here.

Results

A totally of 9 different vocabulary terms were being used either as standalone or in combination with others (connected by logical OR) to define Long COVID across 32 publications. Their counts over years are shown in Figure 1.

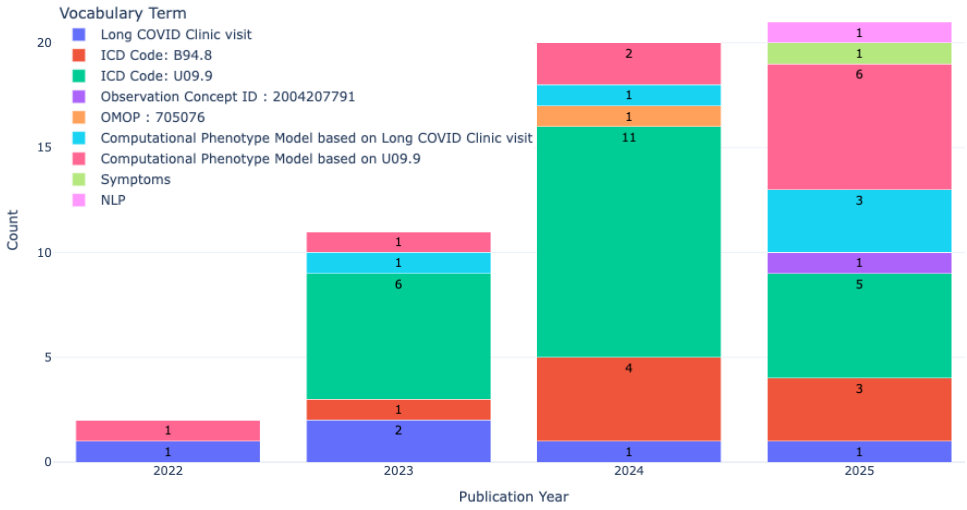

Figure 1: Trend of Vocabulary Terms used to define Long Covid

The results in Figure 1 suggest a few trends. First, human-annotated codes in clinical records such as "ICD-10 code U09.9", "ICD-10 code B94.8", and "Long COVID Clinic Visit" were consistently used to identify Long COVID patients. Second, computational phenotype models (machine learning classifiers) and rule-based methods (NLP rules, symptom set, and OMOP code set) were used to identify patients whose clinical records contained Long COVID symptoms but did not contain U09.9, B94.8, or Long COVID Clinic Visit. Third, the medical informatics community have been employing an increasingly diverse set of terms to define Long COVID over the years.

Discussion

After the initial computational phenotyping model developed by Pfaff et.al,[2] which was initially used to serve as a silver standard to identify Long COVID patients in N3C, clinical practitioners gradually transitioned to using ICD-10 code U09.9 for documenting Long COVID patients. However, the transition was slow and uncoordinated, and the interpretation of symptoms and comorbidities associated with U09.9 was inconsistent across medical institutions. These practical challenges undermined the data quality of U09.9 as a reliable identifier for Long COVID patients, and motivated researchers to use multiple alternative ways to define Long COVID as shown in Figure 1, even after the official definition was published in mid 2024. On the one hand, alternative definitions (computational phenotyping and rule-based methods) demonstrated the community's resilience and creativity in tackling the ambiguity of Long COVID. On the other hand, convergence to an official definition could be accelerated if definitional efforts (such as the NASEM report) that synthesize empirical works and consult clinical experts were prioritized and coordinated in a timely manner. Future work will evaluate the comparative benefits and drawbacks (e.g., reproducibility, sensitivity) of each definitional approach and retrospectively quantify how these differences affect the already published results.

References

1. National Academies of Sciences, Engineering, and Medicine. A Long COVID Definition: A Chronic, Systemic Disease State with Profound Consequences. June 2024. Available from: https://www.nationalacademies.org/read/27768
2. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. Lancet Digit Health. 2022 Jul;4(7):e532–41. doi:10.1016/S2589-7500(22)00048-6 PubMed PMID: 35589549; PubMed Central PMCID: PMC9110014.