Informed Missingness in EHR Labs: A Retrospective Study of Biomarker Trajectories Following SARS-CoV-2 Infection

Saaya Patel[1], Richard Moffitt[1,2]

[1]Biomedical Informatics, Emory University, Atlanta, GA, [2]Hematology and Medical Oncology, Emory University, Atlanta, GA

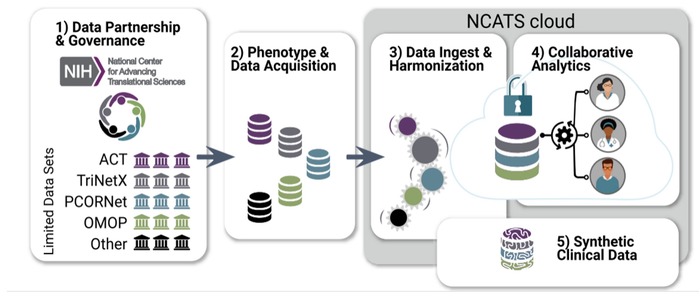
Introduction

- Persistent effects of SARS-CoV-2 infection (Long COVID, LC) remain a major post-pandemic burden, posing diagnostic and management challenges[1]
- Electronic health records (EHRs) enable large-scale study of LC but often contain missing data
- Missingness in EHRs often reflects clinical stability or physician judgment rather than randomness, introducing bias that overrepresents sicker patients[2]
- Incorporating informative missingness directly into analytic models may improve predictive accuracy, interpretability, and generalizability

Objective

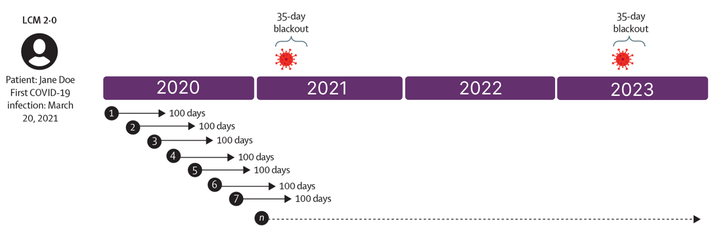
This study evaluates the role of informative missingness in EHR laboratory data, with a case example of lymphocyte trajectories predicting Long COVID.

Methods



Haendel et al., *The National Covid Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment* [3]

Figure 1. Overview of N3C processes and infrastructure



Crosskey et al., *Re-engineering a machine learning phenotype to adapt to the changing COVID-19 landscape*[4]

Figure 2. Long COVID Machine Learning (ML) Model Time Window Selection. Use overlapping 100-day periods to gather data, rather than COVID-19 index date

N3C Database

- Sites: 67
- Persons: 20.6 million
- COVID+ Cases: 8.05 million

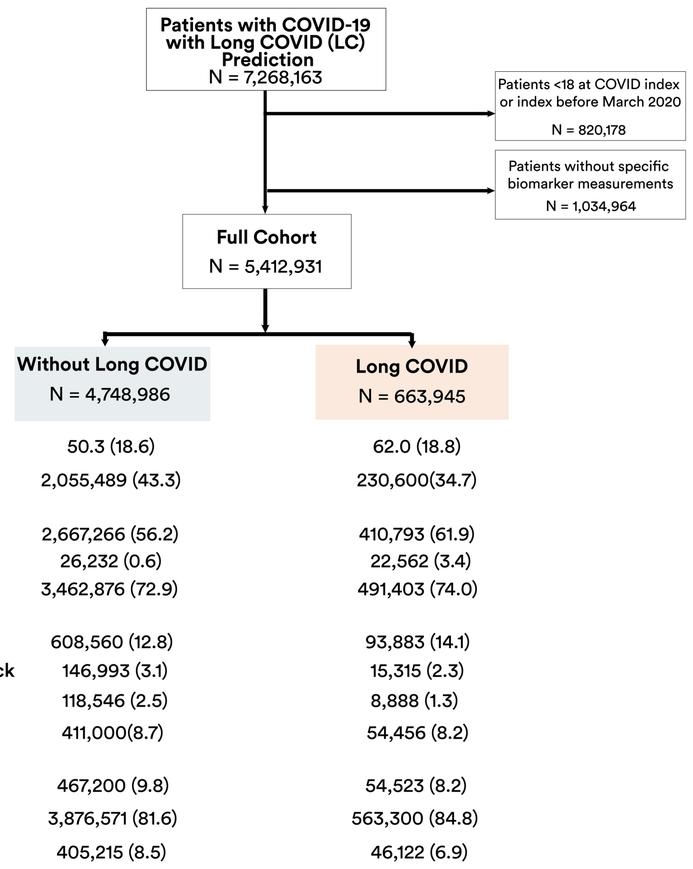
Table 1. Number of sites and patients in N3C database (As of 11/01/2025).

LC Definition

- Training: At least 2 U09.9 ICD Code or LC Clinic Visit
- Features: conditions and observations

Top: Number of outpatients visits, Dyspnoea, Attentiveness

Results



Variables	Without Long COVID N = 4,748,986	Long COVID N = 663,945
Age, Mean (SD)	50.3 (18.6)	62.0 (18.8)
BMI, Mean (SD)	2,055,489 (43.3)	230,600(34.7)
Sex		
Male	2,667,266 (56.2)	410,793 (61.9)
Female	26,232 (0.6)	22,562 (3.4)
Unknown	3,462,876 (72.9)	491,403 (74.0)
Race		
White	608,560 (12.8)	93,883 (14.1)
African American or Black	146,993 (3.1)	15,315 (2.3)
Asian or Pacific Islander	118,546 (2.5)	8,888 (1.3)
Unknown	411,000(8.7)	54,456 (8.2)
Ethnicity		
Hispanic or Latino	467,200 (9.8)	54,523 (8.2)
Non-Hispanic or Latino	3,876,571 (81.6)	563,300 (84.8)
Unknown	405,215 (8.5)	46,122 (6.9)

Figure 3. Study flow diagram of cohort and baseline characteristics of COVID-19 positive patients by Long COVID and hospitalization status.

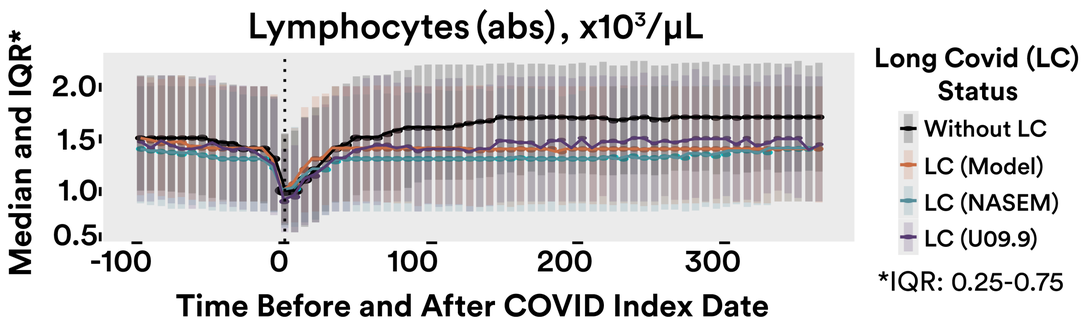


Figure 4. Longitudinal trends by Long COVID status and case definition (Model, NASEM, U09.9, and no LC). Medians (line) and interquartile ranges (bars) shown; time indexed to date of COVID-19 diagnosis (day 0) and spans 100 days before to 360 days after infection.

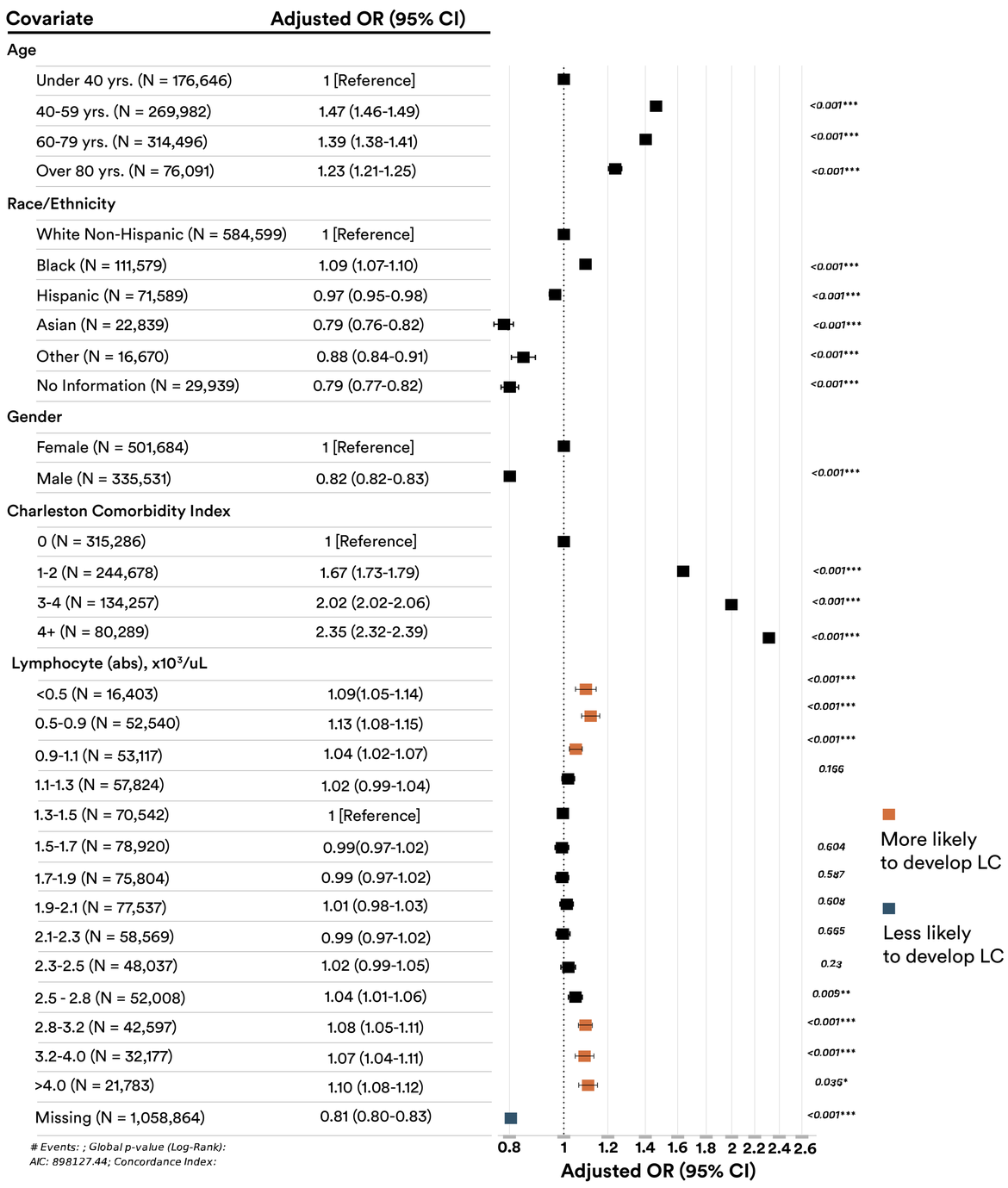


Figure 5. Adjusted odds ratios for long COVID using latest lymphocyte absolute measurement measured 90-180 days post COVID-19 infection. Measurement selected with visit priority to outpatient, inpatient, and intensive care, respectively.

Discussion

- **Lymphocyte Risk Pattern.** Both abnormally low and high lymphocyte counts were linked to increased risk of LC
- **Informative Missingness.** Testing occurred when clinically indicated, producing a Missing Not at Random (MNAR) pattern where absence of testing reflected clinical stability
- **Selection Bias.** As tests were more frequently ordered for patients displaying severe symptoms, potentially skewing analyses that rely exclusively on recorded laboratory data

Future Works

- **Standardize Detection.** Develop reproducible methods to identify and encode informative missingness across multi-institutional EHR datasets

References

1. NHLBI. Long COVID [Internet]. National Institutes of Health (NIH); 2024 Dec. Available from: https://www.nhlbi.nih.gov/covid/long-covid
2. Groenwold RH. Informative missingness in electronic health record systems: the curse of knowing. Diagn Progn Res. 2020;4(1).
3. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association. 2021 Mar 1;28(3):427–43.
4. Bennett TD, Moffitt RA, Hajagos JG, Amor B, Anand A, Bissell MM, et al. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. JAMA Network Open. 2021;4(7):e2116901-1.
5. Crosskey M, McIntee T, Preiss S, Brannock D, Yoo YJ, Hadley E, et al. Reengineering a machine learning phenotype to adapt to the changing COVID-19 landscape: A study from N3C and RECOVER consortia. Lancet Digit Health. 2025.