**Principal component analysis to estimate severity level of Long COVID**

Ângela Jornada Ben, Elizabeth Berghuis-Mutubuki, Johanna Maria van Dongen

**Introduction**

Worldwide, the COVID-19 pandemic has overstretched healthcare systems’ capacity to provide care 1. At the start of the pandemic, people who needed hospitalization and intensive care were those who had priority of care 1. Over the course of the second year, as vaccination rates increased greatly, which in turn reduced the prevalence of serious illness and deaths, the demand for care started to come from those with post-infection complications, also referred to as Long COVID2. Properly identifying such individuals is relevant as it could help to plan prevention and rehabilitation services 3.

Evidence is still limited but suggests that approximately 1 in 10 adults experience ongoing symptoms 3 months after being infected, including those who are asymptomatic at the time of diagnosis3–63–6. The fact that symptoms are heterogeneous and may be different in different populations makes it difficult to accurately identify individuals who are suffering from Long COVID, and thus to assess its burden7.

Recently, a conceptual framework has been proposed by scientists to help define Long COVID at the clinical and pathological levels 7. Nonetheless, consensus does not exist regarding standardized questionnaires that should be used to measure symptoms associated with Long COVID consistently across studies 8. In addition, there is no recommendation on how to properly score the importance of symptoms to define Long COVID and its respective severity levels.

Principal component analysis (PCA) 9, is an unsupervised machine learning technique 10 that has been used to estimate disease severity levels from data that do not contain such information. PCA is typically used to reduce the number of data dimensions by extracting the smallest set of components that may explain most of the variance in the data9. Using PCA, scores can be developed to predict disease severity level based on information available. This method was used to developed severity scores for other chronic diseases such as Multiple Sclerosis and Chronic Obstructive Pulmonary Disease11,12. Therefore, in this study we investigate whether a Long COVID severity level score can be estimated by using a PCA as this information can be used as input to models estimating the burden of Long COVID.

**Methods**

***Data source and settings***

The Dutch Long COVID adult cohort will be used and included participants with a positive PCR test for COVID starting from….and followed util…… Data include symptoms…X

***Long COVID definition***

In this cohort Long COVID was defined as…. By comparing cases with controls…

***Development of the Score by Principal Component Analysis***

Factors related to Long COVID were extracted from the database and used to perform a PCA using the bult-in R function *prcomp*13. We defined the Long COVID severity level as the first principal component (PCA1) score derived by PCA as PCA1 usually is the component that explains most of the variance in the data and, thus, can be associated with high disease severity. The PCA1 score was subsequently estimated based on the factors with high correlation within the PCA1 and using a Radom-effects model, as expressed by the equation:

Where Score(i,t) represents the PCA1 score for each participant (i), indicates the random effect, is the intercept, and represents the error term. Severity levels were classified according to the PCA score percentiles.

**Results**

**Discussion**

**Conclusion**

**References**

1. WHO. COVID-19 continues to disrupt essential health services in 90% of countries [Internet]. 2021 [cited 2022 Feb 5]. Available from: https://www.who.int/news/item/23-04-2021-covid-19-continues-to-disrupt-essential-health-services-in-90-of-countries

2. Ledford H. Do vaccines protect against long COVID? What the data say. Nature. 2021 Nov 23;599(7886):546–8.

3. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of long COVID. Nat Med. 2021 Apr;27(4):626–31.

4. Crook H, Raza S, Nowell J, Young M, Edison P. Long covid—mechanisms, risk factors, and management. BMJ. 2021 Jul 26;n1648.

5. Huang Y, Pinto MD, Borelli JL, Mehrabadi MA, Abrihim H, Dutt N, et al. COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler: Looking for Clarity in the Haze of the Pandemic. medRxiv. 2021 Mar 5;2021.03.03.21252086.

6. NIHR. Living with Covid19 – Second review [Internet]. NIHR Evidence. 2021 [cited 2022 Feb 5]. Available from: https://evidence.nihr.ac.uk/themedreview/living-with-covid19-second-review/

7. Kondratiuk AL, Pillay TD, Kon OM, Lalvani A. A conceptual framework to accelerate the clinical impact of evolving research into long COVID. The Lancet Infectious Diseases. 2021 Jun 1;21(6):756–7.

8. Munblit D, O’Hara ME, Akrami A, Perego E, Olliaro P, Needham DM. Long COVID: aiming for a consensus. The Lancet Respiratory Medicine [Internet]. 2022 May 4 [cited 2022 Jun 7];0(0). Available from: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(22)00135-7/fulltext

9. Jolliffe IT. Principal Component Analysis. Springer Science & Business Media; 2013. 283 p.

10. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT SCI. 2021 Mar 22;2(3):160.

11. Kawachi I, Otaka H, Iwasaki K, Takeshima T, Ueda K. A Principal Component Analysis Approach to Estimate the Disability Status for Patients with Multiple Sclerosis Using Japanese Claims Data. Neurol Ther. 2022 Mar 1;11(1):385–96.

12. Wu EQ, Birnbaum HG, Cifaldi M, Kang Y, Mallet D, Colice G. Development of a COPD severity score. Current Medical Research and Opinion. 2006 Sep 1;22(9):1679–87.

13. Principal Component Analysis in R: prcomp vs princomp - Articles - STHDA [Internet]. [cited 2020 Sep 28]. Available from: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/