cegedim
Health Data

THIN
THE HEALTH IMPROVEMENT NETWORK

LONG COVID SYMPTOMS AND DIAGNOSIS IN PRIMARY CARE: A COHORT STUDY USING THE THIN® DATABASE INCLUDING UNSTRUCTURED TEXT

Anoop D. Shah [1], Samir Dhalla [2], Anuradhaa Subramanian [3], Elizabeth Ford [4], Shamil Haroon [3], Valerie Kuan [1], Krishnarajah Nirantharakumar [3]

[1] Institute of Health Informatics, University College London, London, UK - [2] The Health Improvement Network, a Cegedim Company, Battersea, London, UK
[3] Institute of Applied Health Research, University of Birmingham, Birmingham, UK - [4] Brighton and Sussex Medical School, University of Sussex, Brighton, UK

BACKGROUND

'Long Covid' (post-COVID condition) is a widely recognised consequence of COVID-19 infection, but there is currently little information about the prevalence of symptoms that patients present with in primary care, as these are typically recorded only in free text clinical notes.

OBJECTIVES

To compare the occurrence of symptoms in patients with and without a history of COVID-19 infection, beyond 12 weeks after the infection date.

METHODS

We used data from the The Health Improvement Network (THIN®, a Cegedim Database), a large UK research database of primary care records. We included patients aged 18 or over registered with participating practices in England, Scotland or Wales. We categorised patients into those with confirmed COVID-19 (defined by Read clinical codes for diagnoses or positive test results, or COVID-19 diagnoses extracted from free text), suspected COVID-19, or non-specific viral or respiratory illnesses (some of which may have been COVID-19, as there was little community testing at the time). Suspected or confirmed COVID-19 cases were 1:1 matched to unexposed controls by practice, age, sex, and ethnicity to ensure a similar distribution of key variables and to define an index date for controls, but the matching was not used for analysis. The study period was 1 March 2020 to 31 December 2020.

We extracted baseline and symptom information from both the Read codes and free text, which was analysed using a rule-based named entity recognition and linking algorithm called the Freetext Matching Algorithm (https://github.com/anoopshah/freetext-matching-algorithm). We investigated the incidence of symptom recording relative to the index date (date of COVID-19 infection for cases, or matched time point for controls). We also investigated symptoms recorded within 7 days prior to a GP-recorded confirmed or suspected Long Covid diagnosis (extracted from the free text). We studied 89 symptoms in total, and present results from the 30 most common symptoms.

We calculated hazard ratios for symptom occurrence 12 weeks after the index date adjusted for age, sex, ethnicity, smoking, body mass index and recording of symptoms and consultations prior to the index date, with inverse probability weighting according to a propensity score for being a COVID-19 case based on baseline characteristics and prior diagnoses.

RESULTS

We included 60,800 patients with a mean age of 52 years, of whom 63% were female and 90% were of White ethnicity **(Table 1)**.

	Unexposed controls	Confirmed COVID	Suspected COVID	Viral or respiratory illness
Number of patients	18098	11015	15841	15846
Female, n (%)	11408 (63.0%)	6782 (61.6%)	10093 (63.7%)	10124 (63.9%)
Age in years, mean (SD)	53.6 (18.9)	50.5 (18.1)	52.1 (18.9)	51.5 (19.4)
Ethnicity:				
White	7915 (89.9%)	4589 (89.5%)	6601 (88.9%)	6182 (89.9%)
Black	245 (2.8%)	116 (2.3%)	217 (2.9%)	139 (2.0%)
South Asian	350 (4.0%)	251 (4.9%)	374 (5.0%)	381 (5.5%)
Mixed	81 (0.9%)	46 (0.9%)	98 (1.3%)	51 (0.7%)
Other	214 (2.4%)	126 (2.5%)	138 (1.9%)	122 (1.8%)
Ethnicity not recorded	9293 (51.3%)	5887 (53.4%)	8413 (53.1%)	8971 (56.6%)
Most deprived quintile (at practice level), n (%)	5319 (29.4%)	3602 (32.7%)	4111 (26.0%)	3962 (25.0%)
Median (IQR) number of GP consultations in year prior to index date	5 (2, 10)	6 (2, 12)	9 (4, 16)	8 (4, 15)

Number of days with symptom mentions 1-3 months before index date:				
0	13207 (73.0%)	7273 (66.0%)	8535 (53.9%)	9191 (58.0%)
1	1288 (7.1%)	826 (7.5%)	1382 (8.7%)	1273 (8.0%)
2	1035 (5.7%)	733 (6.7%)	1275 (8.0%)	1155 (7.3%)
3+	2568 (14.2%)	2183 (19.8%)	4649 (29.3%)	4227 (26.7%)
Hospitalised within 28 days of index date	196 (1.1%)	1518 (13.8%)	1887 (11.9%)	902 (5.7%)

Table 1:
Baseline characteristics of patients included in the study
(IQR = interquartile range, SD = standard deviation).

Only 20% of symptom reporting was as Read codes, with 80% in free text. Patients with confirmed COVID-19 had a higher incidence of recording of any symptom for at least 9 months after the index date, compared to a reference period 8-12 weeks before the index date **(Figure 1)**.
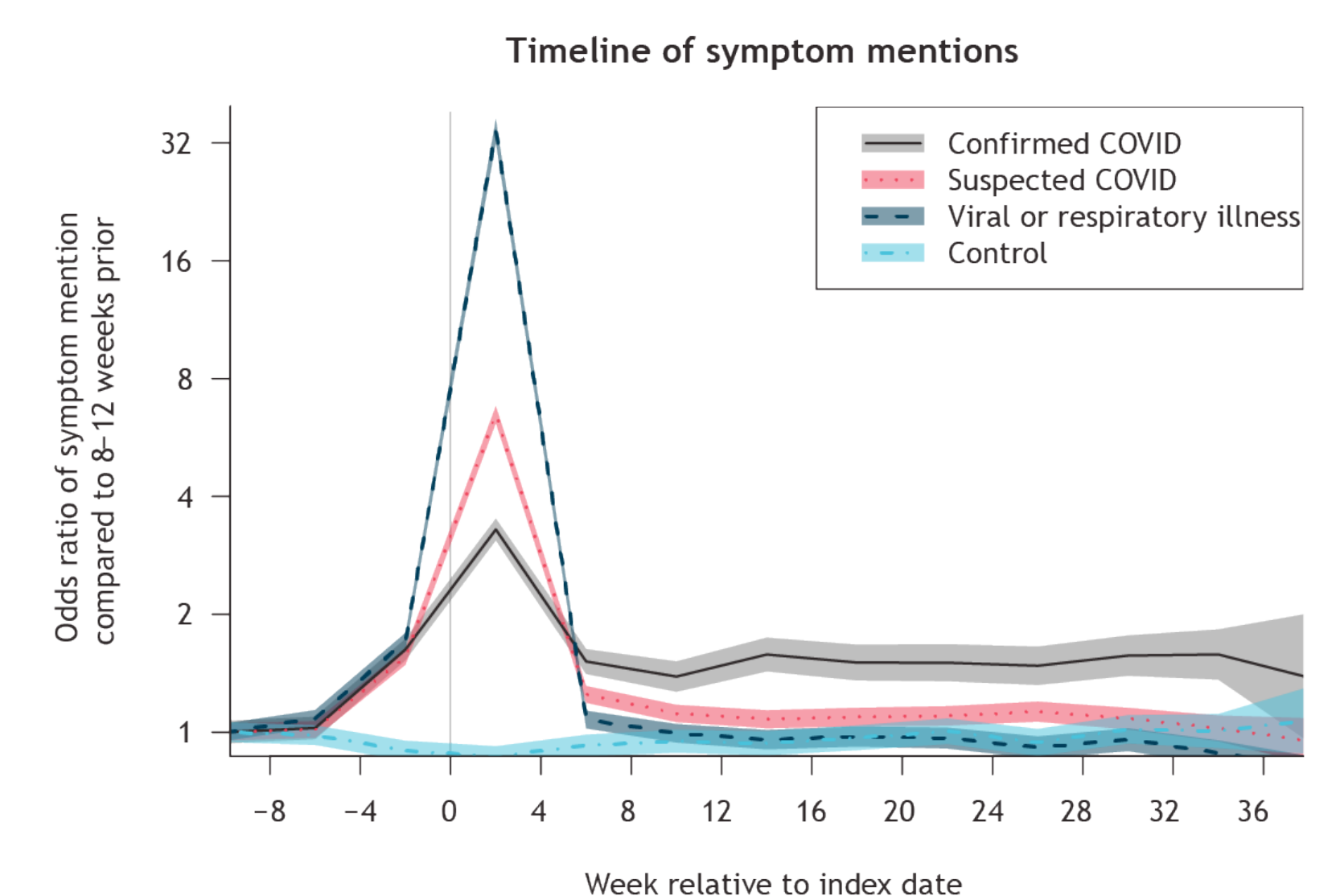

Figure 1:
Timeline of recording of any symptoms by patient category, expressed as odds ratios (estimate and 95% confidence interval) relative to a reference period 8-12 weeks before the index date.

The main analysis compared 11,015 patients with confirmed COVID-19 and 18,098 unexposed controls. A wide range of symptoms were associated with COVID-19 at least 12 weeks previously, with the strongest associations for fatigue (hazard ratio (HR) 4.0, 95% confidence interval (CI) 3.6, 4.4), shortness of breath (HR 3.1, 95% CI 2.9, 3.4), palpitations (HR 2.8, 95% CI 2.3, 3.3), and phlegm (HR 2.9, 95% CI 2.3, 3.6). However, a limited subset of symptoms were recorded within 7 days prior to a Long Covid diagnosis in more than 20% of cases: shortness of breath, chest pain, pain, fatigue, cough, and anxiety / depression **(Figure 2)**.
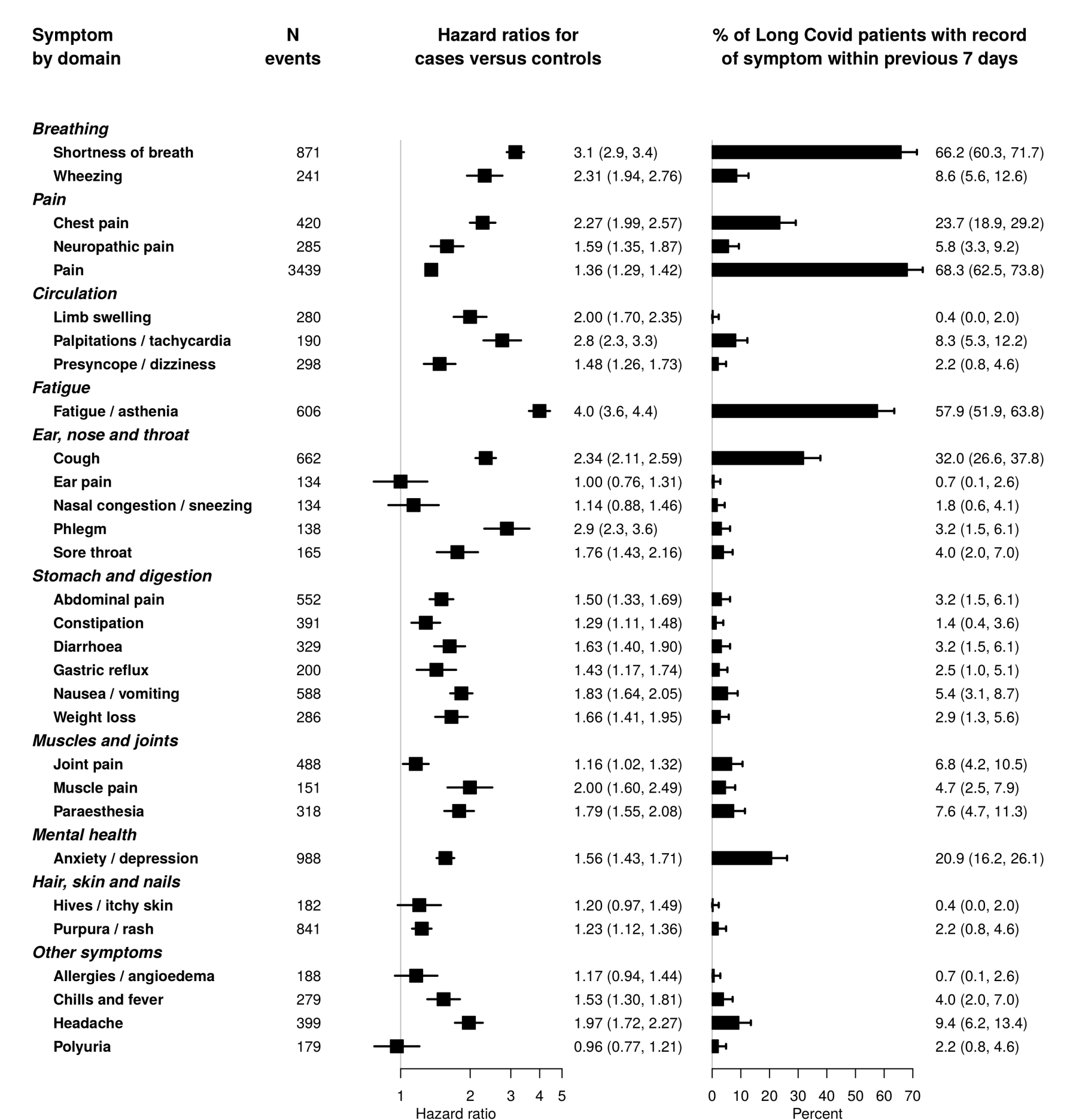

Figure 2:
Adjusted hazard ratios for recording of symptoms at least 12 weeks after the index date, comparing patients with confirmed COVID-19 and unexposed controls, alongside the proportion of patients with a Long Covid diagnosis (at least 12 weeks after the index date) who had the symptom recorded within the preceding week.

CONCLUSIONS

A broad range of symptoms are more commonly reported in patients with a history of COVID-19. However, only a limited subset of these symptoms are typically associated with a GP diagnosis of Long Covid. There is a lack of structured recording of symptoms and COVID-19 diagnoses, showing the importance of free text in health records for studying these topics.