Comparison of methods to identify and characterize Post-COVID syndrome using electronic health records and questionnaires

Bos, Isabelle; Bosman, Lisa; Hoek, Rinske; Waarden, Willemijn; Berends, Matthijs S.; Homburg, Maarten; Olde Hartman, Tim; Muris, Jean; Peters, Lilian; Knottnerus, Bart

DOI:
10.21203/rs.3.rs-3255500/v1

Comparison of methods to identify and characterize Post-COVID syndrome using electronic health records and questionnaires

Isabelle Bos (i.bos@nivel.nl)
Netherlands Institute for Health Services Research

Lisa Bosman
Netherlands Institute for Health Services Research

Rinske Hoek
Netherlands Institute for Health Services Research

Willemijn Waarden
Netherlands Institute for Health Services Research

Matthijs S. Berends
University Medical Center Groningen

Maarten Homburg
University Medical Center Groningen

Tim Olde Hartman
Radboud University Nijmegen Medical Centre

Jean Muris
Maastricht University

Lilian Peters
University Medical Center Groningen

Bart Knottnerus
Netherlands Institute for Health Services Research

Karin Hek
Netherlands Institute for Health Services Research

Robert Verheij
Netherlands Institute for Health Services Research

Research Article

Keywords: post-COVID syndrome, electronic health records, questionnaires, mixed-methods

Posted Date: August 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3255500/v1

Abstract

Background: Some of those infected with coronavirus suffer from post-COVID syndrome (PCS). However, an uniform definition of PCS is lacking, causing uncertainty about the prevalence and nature of this syndrome. We aim to improve understanding by operationalizing different definitions of PCS in different data sources and describing features and clinical subtypes.
Methods: We use different methods and data sources. First, a cohort with electronic health records (EHR) from general practices (GPs) and GP out-of-hours-services combined with sociodemographic data for n≈1.000.000 individuals. Second, questionnaires among n=276 individuals who had been infected with coronavirus. Using both data sources, we operationalized definitions of PCS to calculate frequency and characteristics. In a subgroup of the EHR data we conducted community detection analyses to explore possible clinical subtypes of PCS.

Results: The frequency of PCS ranged from 15-33%, depending on the method and data source. Across all methods and definitions, the mean age of individuals with PCS was around 53 years and they were more often female. There were small sex differences in the type of symptoms and overall symptoms were persistent for 6 months. Exploratory network analysis revealed three possible clinical subtypes.

Discussion: We showed that frequency rates of post-COVID syndrome differ between methods and data sources, but characteristics of the affected individuals are quite stable. Overall, PCS is a heterogeneous syndrome affecting a significant group of individuals who need adequate care. Future studies should focus on care trajectories and qualitative measures such as experiences and quality of life of individuals living with PCS.

INTRODUCTION

Already in the first year of the global corona pandemic it became clear that a substantial part of the individuals infected with the coronavirus (SARS-CoV-2) suffer from persistent symptoms, also called 'Post-COVID-syndrome' (PCS) [1]. Although literature about PCS - or other related terminology like 'Long COVID' and 'Post-acute sequalae of COVID-19'- is now growing rapidly [2]–[6], there is still much debate about the prevalence and the characteristics of people affected by PCS which is partly due to a lack of uniform definition and is believed to strongly depend on the data sources and methods used in the conducted research. In addition, there have been some studies which aimed to identify different phenotypes or clinical subtypes of PCS [7], [8], but further research and validation is required as results regarding subtypes point in different directions thus far. Identification and agreement regarding clinical subtypes of PCS would aid to optimize the diagnostic process, provide clarity about the prognosis and personalize treatment plans.

Prevalence estimates reported so far on PCS vary widely between studies ranging from 13–80% [5], [6], [9], [10]. The wide variety in reported prevalence estimates is at least partly due to variations in the investigated population (e.g. hospitalized individuals due to the corona virus only vs. nationwide or vaccinated vs. non-vaccinated) and to the definition of PCS, which has been described in various ways. Like many definitions, the definition of the World Health Organization (WHO) leaves room for interpretation and variation, as it describes PCS as a condition in which individuals have persistent symptoms three months after the onset of COVID-19 infection that last for at least three months that cannot be explained by an alternative diagnosis [11]. In this definition there is no clarity about the most important symptoms included, leaving ample room for heterogeneity amongst studies and in clinical practice. Moreover, many studies are solely based on self-reported symptoms and lack the ability to compare to control groups without persistent symptoms or without a COVID-19 infection, leaving them unable to identify specific PCS symptoms and associations with for instance demographic factors.

In this study, we examine the impact of using different definitions and data sources on the estimated frequency of PCS and the characteristics of individuals suffering from it. Moreover, we aim to examine clinical whether we can identify clinical subtypes of PCS using EHR data and machine learning. Defining the population affected and exploring the existence of clinical subtypes of PCS are necessary steps, to carry out subsequent analyses of care trajectories of the patients involved and possible predisposing factors related to PCS. We will make use of a large database with primary care electronic health records (EHR) and a series of questionnaires collected over a years’ time among individuals infected with COVID-19. With these data sources we aim to provide a clear definition of this syndrome, identify characteristics of the individuals suffering from PCS and explore whether we can identify clinical subtypes of PCS. The results of this study will be used for follow up studies describing and comparing different care trajectories of individuals with PCS and to focus on specific aspects of PCS like quality of life.

METHODS

The current study is part of the *Long COVID MM* (Long COVID Mixed Methods) project in which various methods are combined to provide insight into post-COVID-syndrome. This project was conducted according to the Declaration of Helsinki and ethical approval was obtained from the medical ethics committee (METc) from the VU University Medical center Amsterdam for the longitudinal questionnaire component (METc protocol number 2020.0709) and from the METc of the University Medical Center Groningen for the electronic health records component (METc protocol number 2021/473).

Data sources

GP-EHR cohort

Data of this cohort consists of electronic health records (EHR) from general practitioners (GP) and GP out-of-hours-services (OOH-services) combined on individual-level with demographic and socio-economic data. The EHR data from GPs and GP out-of-hours-services is obtained via Nivel Primary Care Data base (Nivel-PCD; approved under number NZR-00321.052) which uses an opt-out system permitted under the Dutch Medical Treatment Contracts Act (WGBO). The GP-EHR data include age, sex, prescriptions (coded via Anatomical Therapeutic Chemical classification, ATC), contacts, referrals, lab results and diagnosis or symptoms (International Classification of Primary Care-1, ICPC-1, coded). The OOH-services data include: contacts with diagnosis or symptoms (ICPC-1 coded), prescriptions (ATC-coded) and triage registration (ICPC-1 coded). All EHR data is pseudonymized and linked on individual-level by a trusted third party. The GP-EHR data is stored and combined with demographic and socio-economic data on the data platform from Statistics Netherlands, including: age, sex, migration background, education level, household income and mortality data (date and cause of death). For the current study the GP-EHR cohort has combined data from Nivel-PCD and Statistics Netherlands 2019 and 2020 for n = 958,739 individuals in total.

Corona Survey Cohort

As described elsewhere [12], this population-based cohort was initiated in 2020 to study the long term effects of COVID-19 infection and is based within Nivel-PCD. As part of the Long COVID MM study the initial cohort was extended with more participants and an additional follow-up survey. In short, n = 1851 individuals who had been flagged in their electronic patient file as having COVID-19 by their GP (ICPC code R83.03) were invited to participate between January and September 2021 in this study by their general practitioner (GP). Individuals were sent four surveys: direct after inclusion and after 3, 6 and 12 months. The surveys contained questions on symptoms, used health care and care experiences, quality of life, ability to work, vaccination status and selfcare. All participants signed informed consent forms allowing researchers to link the survey data to their GP EHR. This enables the unique opportunity to combine the survey data with EHR data on morbidities and prescriptions for this specific group. More details about the Corona Survey Cohort are described in a previous publication [12].

Data analysis

Definition Post-COVID-syndrome: GP-EHR Cohort

In the GP-EHR cohort we conducted the following analyses to explore manners to operationalize different PCS definitions using routine healthcare data. First we selected individuals with COVID-19 (n = 10,313) based on the EHR data from Nivel-PCD who had been flagged as having COVID-19 by their GPs directly, or who had been identified via a developed algorithm selecting patients based on symptoms and episode titles between April and June 2020 [13]. Each individual with COVID-19 was matched to four control individuals without COVID-19. Matching to control patients was only used for the operationalization of the definition. Controls were similar to individuals with COVID-19 in age and sex and were followed over the same period in the data as the individuals with COVID-19 to adjust for seasonal or circumstantial effects like lockdowns due to COVID-19. Characteristics of the matched control group can be found in Supplementary Table 1. To create a list of symptoms related to PCS, which is needed for the definition of PCS, we compared the ICPC codes recorded in the COVID-19 group, 3–12 months after infection, to the ICPC codes recorded in the same individuals a year before infection and to the ICPC codes recorded in the control group. Similar to the WHO definition we choose 3 months after COVID-19 infection as the cut-off point between acute COVID-19 symptoms and PCS [11]. We created symptom lists following several steps. First, we ranked the ICPC codes by prevalence in the COVID-19 group nine months before and 3–12 months after infection. Thereafter, we calculated the difference in prevalence before and after infection and created a top 30 list of ICPC codes of which the prevalence was increased after infection. We excluded ICPC codes that had also increased in the matched control group. This list was reviewed by four GP-researchers to exclude symptoms that were unlikely to be related to PCS. A total of n = 25 symptoms were included in the 'data-derived list'. In addition, we compiled a list of symptoms published by the WHO [11] and expanded this with symptoms reported by participants of the Corona Survey Cohort and a panel of 8 patients (age range: 32–75 years, 4 males and 4 females) who provide advice and feedback during the project. This 'patient reported list' included a total number of n = 37 symptoms. Furthermore, a GP (MH) and a medical microbiologist (MB) independently reviewed the entire list of ICPC codes for symptoms that could be related to acute COVID-19 and possibly also to PCS. This 'clinicians (acute) COVID list' included n = 30 possible symptoms. We compared the symptoms on these three lists (i.e. 'data-derived list', 'patient reported list' and 'clinicians (acute) COVID list') and symptoms that were included on at least two lists were considered 'core symptoms', while the remaining symptoms were considered 'additional symptoms' (Supplementary Table 2). We used the core and additional symptoms to classify the individuals with COVID-19 as having PCS using their EHR data according to a broad and narrow definition. According to the broad definition patients should have consulted the GP for at least one core symptom or at least two different additional symptoms, 3–12 months after COVID-19 infection. According to the narrow definition patients should have consulted the GP for at least two symptoms of which minimal one core symptom and at least two consultations for these symptoms at the GP. These definitions were created based on current literature regarding PCS definitions (e.g. but not limited to [11], [14], [15], input from GPs (JM, TOH, BK) and the involved researchers. We developed these two definitions because there is currently no uniform definition and we aimed to investigate the influence of using different definitions. The broad definition was created to be inclusive to possible heterogeneity of the syndrome and the narrow definition to depend more on care usage and the core symptoms.

Definition Post-COVID-syndrome: Corona Survey Cohort

Of the total number of participants in the Corona Survey Cohort (n = 442), n = 276 (62%) participants were selected for the current analysis as they completed the first questionnaire within 3 months after COVID-19 infection (between January – September 2021) and could answer the questions regarding acute symptoms more accurately. Individuals were classified as having PCS when they reported at least one symptom, from a selected list of symptoms, three months after the COVID-19 infection and experienced discomfort in their daily living (first survey) or reported not be recovered after the initial COVID-19 infection (second survey). Individuals were classified as non-PCS (n = 93) when they reported not to experience discomfort or reported to be recovered. Individuals were classified as 'unknown' (n = 92) when relevant data to classify individuals as PCS or non-PCS was missing or when there was a discrepancy in the answers to the questionnaire (i.e. report no symptoms, but also report not to be recovered).

Classifications and statistical analysis

We used descriptive statistics to describe the sample characteristics of the PCS patients in the combined EHR cohort and the Corona Survey Cohort. Age categories were divided into: children and adolescents (age 0–23 years of age), adults (23–70 years of age), elderly (≥ 70 years of age). Migration background was dichotomized as: both parents were born in the Netherlands (0) and at least one parent is not born in the Netherlands (1). Education level was divided into low (primary school of pre-vocational education), medium (secondary or vocational education) and high (professional higher education or university) education level. Income level was only available in the GP-EHR cohort and was divided according to standardized household income in the Netherlands into low (0–40 percentile), medium (40–80 percentile) and high (> 80 percentile). GP consultations were defined as long, medium and short consultations including consultations by phone and email and long and short visitations. Long and short consultations with the nurse practitioner were also included. Additionally in a subgroup of the EHR cohort we performed a network analysis, Louvain Community Detection [16], to identify symptoms that often co-occur in individuals with PCS. For these analyses we only included individuals who consulted the GP with more than one symptom (n = 1,503) and we used the R-package 'igraph' for visualization. A community was included in the network when at least 1% of individuals have this combination of symptoms and we used a cut-off of > 0.3 on the modularity score to ensure the quality of the communities and network [17]. We then classified the 1503 individuals into individuals with at least two symptoms in one community and described the demographic characteristics of these individuals. Statistical analyses were performed in STATA (version 16.1) and R (version 4.1.3).

RESULTS

The selection of the GP-EHR cohort we used for this analysis includes data on n=10,313 individuals who were all infected with the COVID-19 virus. Of these individuals, n=452 (4.3%) were hospitalized due to COVID during the acute COVID phase (0-3 months after infection). Table 1 describes the characteristics of these individuals, classified according to the broad and narrow definition of PCS. The selection of the Corona Survey Cohort we used for the current analysis included n=276 individuals who had been infected by COVID-19. Of these individuals n=18 (6.6%) had been hospitalized during the acute COVID phase (0-3 months after infection). Table 2 describes the characteristics of the individuals from the Corona Survey Cohort. The percentages of individuals classified as having PCS ranged from 15-33% depending on the definition and data source (Table 1 & Table 2).

[insert Table 1 here]

GP-EHR cohort: Demographics and other characteristics using broad and narrow PCS definitions

In the GP-EHR cohort comparisons were made between individuals in the PCS group and the non-PCS group according to the broad and narrow definitions (Table 1). Results of the comparisons were similar for both definitions and therefore we only mention the results using the narrow definition in the text. Individuals with PCS were more often female (69% vs. 57%, p≤0.001) and were older (53.4 vs. 51.1 years, p≤0.001) compared to the non-PCS group There were significantly fewer children in the PCS group compared to the non-PCS group and more adults and elderly in the PCS groups (Table 1). There was no difference between the PCS group and the non-PCS group in education level, household income or migration background. The average number of symptoms for which the GP was consulted by individuals with PCS was 6.8 (SD 5.4) symptoms per patient versus 0.9 (SD 1.8) in the non-PCS group ($p \leq 0.001$). The average number of GP consultations was, by definition, higher in the PCS group compared to the non-PCS group (5.5 vs. 0.8 consultations, $p \leq 0.001$).

[insert Table 2 here]

Corona Survey Cohort: Demographics and other characteristics in PCS and non-PCS group

In the Corona Survey Cohort (Table 2) we compared the PCS group to the non-PCS group. The characteristics of the unknown group (n=92) can be found in Supplementary Table 3. Unlike in the GP-EHR cohort, there was no significant difference in age and sex between the PCS group and the non-PCS group. Individuals in the PCS group more often had a lower education level ($p \leq 0.001$) compared to the non-PCS group. We compared the average number of self-reported symptoms in the PCS group to the non-PCS group 3 (9.2 vs. 2.8; $p \leq 0.001$) and 6 months (7.2 vs. 2.1; $p \leq 0.001$) after infection. Twenty-one (23%) individuals with PCS reported that they are working less or stopped working due to PCS symptoms after 3 months and 11 (16%) after 6 months (Table 2).

[insert Figure 1 here]

GP-EHR cohort: Frequency of symptoms stratified by sex

Figure 1 shows the frequencies of patients that visit the GP for a particular symptom 3-12 months after the COVID-19 infection stratified by PCS definition and sex for the top 10 most prevalent core symptoms. For all symptoms we found that males consulted their GP less often for these symptoms compared to females. The most prevalent symptoms in females were psychological symptoms (22-25%) including anxiety and depression, while respiratory symptoms (15-19%) like coughing or dyspnea were most prevalent in males.

[insert Figure 2 here]

Corona Survey Cohort: Frequency of symptoms over time in males and females

Figure 2 shows the frequency of symptoms for the PCS patients in the Corona Survey Cohort at 3 and 6 months, stratified by gender. Overall, symptom frequencies are considerably higher in the Corona Survey Cohort compared to the GP-EHR cohort and different symptoms are reported (Figure 1). In the Corona Survey Cohort, the most prevalent and persistent symptom was fatigue in both males (3 months: 89%, 6 months 78%) and females (3 months: 89%, 6 months 86%). Similar to the GP-EHR cohort, males reported less symptoms than females and males more often reported respiratory symptoms compared to females at 3 months (58% vs. 45%), while this is opposite at 6 months (44% vs. 51%). Most symptoms were reported at the same frequency or increase over time (3 to 6 months), except for excessive sweating and loss of smell and/or taste which seems to decrease over time.

[insert Figure 3 & Table 3 here]

GP-EHR cohort: Community detection analyses to explore clinical subtypes

To explore possible clinical subtypes of PCS community detection analyses were performed in a subgroup of individuals in the GP-EHR cohort (n=1,503) who visited the GP for at least two different symptoms. Figure 3 shows the results of the community detection of the combination of symptoms that often occur together. We identified a network with a modularity score (possible range: -0.5 to 1.0) of 0.302 indicating a network with average strength in which three communities with symptoms were identified (Figure 3). Community A includes psychological and generalized symptoms and was statistically significant (p=0.045), Community B includes cardiorespiratory symptoms (p=0.494) and Community C includes gastrointestinal symptoms (p=0.617). The communities are solely based on symptoms that co-occur and not on how many individuals have only these combinations of symptoms. Therefore, we subsequently analyzed how many individuals of this subgroup (n=1,503) could be classified as experiencing this combination of symptoms (i.e. having at least two symptoms within one community). When classifying the group into individuals with at least two ‘community symptoms’ we found that n=248 (17%) had symptoms across communities and n=458 (30%) had symptoms that were not included in the network. In addition, there were individuals with only a single ‘community symptom’, n=360 (24%) in community A, n=126 (8%) in community B and n=150 (10%) in community C. Table 3 shows the characteristics of the individuals who could be classified as experiencing distinct community symptoms. The group with neuro-respiratory symptoms (Community A) was the largest group (n=109, 7%) and often females with an average age of 54.2. Thirty-two (2%) individuals experienced only symptoms from community B which are gastrointestinal symptoms. These individuals were younger (mean age 51.2) and included the lowest percentages of females (68%). The group with cardiopulmonary symptoms was the smallest including n=18 (1%) who had an average age of 55.3 years. The percentages of individuals with a migration background was similar across the community groups (22-24%; Table 3).

DISCUSSION

This study describes the extent to which definitions and data collection methods affect the frequency of post-COVID syndrome as well as its constituting symptoms. By combining the results of the analyses of GP-EHR data from more than 50.000 individuals and longitudinal questionnaire in over 275 individuals the main findings are: 1) the frequency of PCS among individuals infected by the coronavirus between April – July 2020 in the Netherlands, ranged from 15–33% depending on the definition and data source used; 2) individuals with PCS were on average 53 years old and more often female; 3) individuals with PCS consulted the GP most often with psychological problems while fatigue was the most often self-reported symptom; 4) three communities of related PCS symptoms could be identified but require further examination and validation to define clinical subtypes of PCS.

Thus far worldwide prevalence rates of PCS vary widely between studies depending on populations, methods and definitions. Two recent meta-analysis regarding the prevalence of PCS for instance showed that prevalence rates were higher among individuals who were hospitalized during the acute phase compared nonhospitalized populations [14], [15]. Our study includes both nonhospitalized and hospitalized individuals in the GP-EHR cohort (4% hospitalized) as well as in the Corona Survey cohort (7% hospitalized), although our group of hospitalized individuals are small in comparison to hospitalization rates due to COVID-19 in the Netherlands at that time [18], [19]. This might cause for a slight underestimation of the PCS frequency in our study. Another important factor influencing the variety among prevalence rates reported is whether a control group and comparisons on individual level with pre-COVID situation regarding symptoms and comorbidities has been included [15]. The few studies that have also included this crucial correction for a control group, like our study, generally report lower prevalence rates, similar to our results [6], [20]. Other obvious but central differences between studies on prevalence rates are whether PCS is defined based on self-report and whether patients are included based on only one symptom or on multiple symptoms. Our results underline and clarify the influence of these factors and the impact it has on the characterization of the group individuals suffering from PCS as we compared a broad (minimal 1 symptom) and narrow (minimal 2 symptoms and multiple consultations) definition in the EHR data and the self-report data from the Corona Survey Cohort. Besides the obvious influence the narrow and broad definition have on the size of the PCS group it did not influence the characteristics (i.e. age, gender, migration background) of the individuals included. On the other hand, when comparing the PCS group in self-report survey data (Corona Survey Cohort) to EHR data we did find a noteworthy difference in the level of education between the PCS and the non-PCS group which we did not find in the EHR data. In the survey data we found a higher percentage of individuals with a low education in the PCS group compared to the non-PCS group and the total group. This finding is in line with a German study which also showed that higher level of education was associated with a lower risk of PCS [21]. The lack of association in the GP-EHR cohort could be due to the large number of individuals for whom the level of education was unknown (38%), although the distribution among the education categories in the group of individuals in the PCS group for whom this is known (62%) is similar to the distribution in the total COVID group. Future studies should further investigate the association between PCS and education level to validate our findings.

In general, findings thus far published regarding sex and PCS are quite consistent and also in line with our results as most studies report a higher occurrence of PCS in females compared to males [3], [14], [22], [23]. In addition, we also found that females report or seek help for different PCS symptoms than males. Females with PCS more often consult the GP for mental health symptoms, while males consult the GP most often for respiratory symptoms. A previous study also reported sex differences in relation to PCS symptoms but only included somatic symptoms and no psychological or mental symptoms [6]. Nevertheless, our results are in line with a large body of literature showing that males are less likely to seek medical help, in particular for mental health problems (Galdas, Cheater, and Marshall 2005). In general the most prevalent PCS symptoms for which the GP is consulted are psychological symptoms including anxiety and depressed mood, digestive symptoms including diarrhea and obstipation and respiratory symptoms including dyspnea and trouble breathing. Surprisingly, fatigue is not the most often reported symptom at the GP while a meta-analysis reported that this is the most common symptom of

PCS [25]. Yet when focusing on self-reported symptoms in our survey data, fatigue is found to be the most common symptom. This emphasizes the differences between using routine healthcare registry data and self-report data which has been reported before [26]–[28], but requires further examination in relation to PCS.

To examine whether we could use machine learning to identify specific clusters of symptoms that often co-occur we performed a community detection analysis. We identified three communities which only partly overlap with clusters identified in a previous study which identified clusters across different SARS-CoV-2 variants [8]. Similar to our findings, Canas and colleagues (2022) also identified a cluster with mainly cardiorespiratory symptoms which was associated with the wild-type variant of the virus (i.e. first stages of the pandemic). The other clusters and communities associated with the early variant of the virus were however different from our findings [8]. To our knowledge there are no other studies using a data-driven approach to identify clinical subtypes of PCS. Literature studies on symptoms that often co-occur do however also often show a cardiorespiratory cluster [29], a generalized-mood cluster [30] and a gastrointestinal cluster [31]. Nevertheless, not all literature points in the same direction as also other clusters have been mentioned depending on population, virus variant and clustering method [32]. Our results are not conclusive on the clinical subtypes and should be interpreted with caution as groups were small and there were many individuals with symptoms in multiple clusters or other combinations of symptoms not identified by the analyses. It is also important to mention that the community detection analyses were conducted in a subgroup of individuals whom consulted the GP for various symptoms, which could indicate a more severe phenotype of PCS and might not occur in all individuals with PCS. Future studies, perhaps using data with biological and continuous parameters, should validate and further examine possible phenotypes of PCS as EHR might be not be well suited for these types of analysis due to the categorical coding and registration limitations [33].

The major strengths of this study are the large sample size and the generalizability of the GP-EHR cohort, including the ability to compare to reference groups (non-PCS and non-COVID) and data from before the COVID pandemic. In addition, the combination of methods (using EHR data and surveys) allows for internal validation and interpretation of the results and provides a unique opportunity to compare frequency rates and symptoms reported in self-report data and routine healthcare data. Also, by using different sources of input (data-driven, list of WHO and experts) to create a list with core and additional symptoms added rigor to our study. However, also some limitations of this study should also be acknowledged. First, we identified patients as having had a COVID-19 infection when they visited their GP with COVID related symptoms and not by including all patients who were tested positive for COVID-19 by the national testing authorities as public testing was not yet available during this time period. Second, in the survey data (Corona Survey Cohort) there may be a selection bias as individuals who are experiencing persistent symptoms may be more likely to complete questionnaires compared to individuals not experiencing symptoms. Third, in this paper we only focused on individuals who were infected with COVID-19 in the first period of the COVID-19 pandemic and therefore not all variants of the virus are included. In future studies it would be possible to examine the relationship between virus variants and PCS.

In conclusion, our results indicate how definitions and the choice of data sources affect the frequency of PCS and the characteristics of the individuals affected by it as well as symptoms that are regarded as part of it. Frequency rates differ between methods and data sources (15–33%) but characteristics of the affected individuals seem less affected as we found that PCS mostly affects middle-aged females. The insights from this study form a solid basis for subsequent analyses on quality of life, care trajectories and risk factors for developing PCS. These analyses will be conducted in the near future to improve understanding and care for individuals with PCS, which is desperately needed.

Declarations

Ethical Approval

This project was conducted according to the Declaration of Helsinki and ethical approval was obtained from the medical ethics committee (METc) from the VU University Medical center Amsterdam for the longitudinal questionnaire component (METc protocol number 2020.0709) and from the METc of the University Medical Center Groningen for the electronic health records component (METc protocol number 2021/473). Conditions are fulfilled under which the use of electronic health records for research purposes in the Netherlands is allowed. Under these conditions, neither informed consent from study subjects nor approval by a medical ethics committee is obligatory for this type of observational studies, containing no directly identifiable data (art. 24 GDPR Implementation Act jo art. 9.2 sub j GDPR). All participants of the Corona Survey Cohort gave informed consent before starting the survey and could additionally provide informed consent for linkage of EHR data to questionnaire data.

Availability of data and materials

The data that support the findings of this study are available upon request via the corresponding author Nivel but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Consent for publication

Not applicable.

Competing interests

All authors declare no competing interests.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analyses were performed by IB, LB, RvdH, WvW and KH. The first draft of the manuscript was written by IB and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This study was performed as part of the *Long COVID MM* project which is financed by ZonMw under project number 10430302110004.

References

1. A. Carfì, R. Bernabei, F. Landi, and Gemelli Against COVID-19 Post-Acute Care Study Group, 'Persistent Symptoms in Patients After Acute COVID-19', *JAMA*, vol. 324, no. 6, pp. 603–605, Aug. 2020, doi: 10.1001/jama.2020.12603.
2. R. S. Peter *et al.*, 'Prevalence, determinants, and impact on general health and working capacity of post-acute sequelae of COVID-19 six to 12 months after infection: a population-based retrospective cohort study from southern Germany'. medRxiv, p. 2022.03.14.22272316, Mar. 15, 2022. doi: 10.1101/2022.03.14.22272316.
3. M. Taquet, Q. Dercon, S. Luciano, J. R. Geddes, M. Husain, and P. J. Harrison, 'Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19', PLoS Med., vol. 18, no. 9, p. e1003773, Sep. 2021, doi: 10.1371/journal.pmed.1003773.
4. B. Meza-Torres *et al.*, 'Differences in Clinical Presentation With Long COVID After Community and Hospital Infection and Associations With All-Cause Mortality: English Sentinel Network Database Study', JMIR Public Health Surveill., vol. 8, no. 8, p. e37668, Aug. 2022, doi: 10.2196/37668.
5. S. Lopez-Leon *et al.*, 'More than 50 Long-term effects of COVID-19: a systematic review and meta-analysis', *MedRxiv Prepr. Serv. Health Sci.*, p. 2021.01.27.21250617, Jan. 2021, doi: 10.1101/2021.01.27.21250617.
6. A. V. Ballering, S. K. R. van Zon, T. C. olde Hartman, and J. G. M. Rosmalen, 'Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study', *The Lancet*, vol. 400, no. 10350, pp. 452–461, Aug. 2022, doi: 10.1016/S0140-6736(22)01214-4.
7. S. J. Yong and S. Liu, 'Proposed subtypes of post-COVID-19 syndrome (or long-COVID) and their respective potential therapies', Rev. Med. Virol., vol. 32, no. 4, p. e2315, 2022, doi: 10.1002/rmv.2315.
8. L. S. Canas *et al.*, 'Profiling post-COVID syndrome across different variants of SARS-CoV-2', Health Informatics, preprint, Jul. 2022. doi: 10.1101/2022.07.28.22278159.
9. C. Fernández-de-las-Peñas *et al.*, 'Prevalence of post-COVID-19 symptoms in hospitalized and non-hospitalized COVID-19 survivors: A systematic review and meta-analysis', *Eur. J. Intern. Med.*, vol. 92, pp. 55–70, Oct. 2021, doi: 10.1016/j.ejim.2021.06.009.
10. C. H. Sudre *et al.*, 'Attributes and predictors of long COVID', *Nat. Med.*, vol. 27, no. 4, Art. no. 4, Apr. 2021, doi: 10.1038/s41591-021-01292-y.
11. WHO, 'A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021', 2021. https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1 (accessed Jul. 08, 2022).

12. R. Veldkamp, K. Hek, R. van den Hoek, L. Schackmann, E. van Puijenbroek, and L. van Dijk, ‘Combining General Practice Electronic Records with Patient Reported Outcomes to study impact of COVID-19 in patients in the general population: a description of the Nivel Corona Cohort’, Submitted.
13. M. Hooiveld *et al.*, ‘Weekcijfers COVID-19-patiënten in de huisartsenpraktijk. Week 10–27, 2 maart – 5 juli 2020. | Nivel’, Jul. 09, 2020. https://www.nivel.nl/nl/publicatie/weekcijfers-covid-19-patienten-de-huisartsenpraktijk-week-10-27-2-maart-5-juli-2020 (accessed Aug. 11, 2022).
14. C. Chen, S. R. Haupert, L. Zimmermann, X. Shi, L. G. Fritsche, and B. Mukherjee, ‘Global Prevalence of Post COVID-19 Condition or Long COVID: A Meta-Analysis and Systematic Review’, J. Infect. Dis., p. jiac136, Apr. 2022, doi: 10.1093/infdis/jiac136.
15. L. L. O’Mahoney *et al.*, ‘The prevalence and long-term health effects of Long Covid among hospitalised and non-hospitalised populations: A systematic review and meta-analysis’, *eClinicalMedicine*, vol. 55, p. 101762, Jan. 2023, doi: 10.1016/j.eclinm.2022.101762.
16. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, ‘Fast unfolding of communities in large networks’, *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.
17. J. Xie and B. K. Szymanski, ‘Community Detection Using A Neighborhood Strength Driven Label Propagation Algorithm’, in 2011 *IEEE Network Science Workshop*, Jun. 2011, pp. 188–195. doi: 10.1109/NSW.2011.6004645.
18. L. Coyer *et al.*, ‘Hospitalisation rates differed by city district and ethnicity during the first wave of COVID-19 in Amsterdam, The Netherlands’, BMC Public Health, vol. 21, no. 1, p. 1721, Sep. 2021, doi: 10.1186/s12889-021-11782-w.
19. E. Mathieu *et al.*, ‘Coronavirus Pandemic (COVID-19)’, *Our World Data*, Mar. 2020, Accessed: Dec. 16, 2022. [Online]. Available: https://ourworldindata.org/covid-cases
20. A. B. C. Cazé *et al.*, ‘Prevalence and risk factors for long COVID after mild disease: a longitudinal study with a symptomatic control group’. medRxiv, p. 2022.09.15.22279958, Sep. 22, 2022. doi: 10.1101/2022.09.15.22279958.
21. T. Bahmer *et al.*, ‘Severity, predictors and clinical correlates of Post-COVID syndrome (PCS) in Germany: A prospective, multi-centre, population-based cohort study’, *eClinicalMedicine*, vol. 51, p. 101549, Sep. 2022, doi: 10.1016/j.eclinm.2022.101549.
22. P. Montenegro *et al.*, ‘Prevalence of Post COVID-19 Condition in Primary Care: A Cross Sectional Study’, *Int. J. Environ. Res. Public. Health*, vol. 19, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/ijerph19031836.
23. F. Bai *et al.*, ‘Female gender is associated with long COVID syndrome: a prospective cohort study’, *Clin. Microbiol. Infect.*, vol. 28, no. 4, p. 611.e9-611.e16, Apr. 2022, doi: 10.1016/j.cmi.2021.11.002.
24. P. M. Galdas, F. Cheater, and P. Marshall, ‘Men and health help-seeking behaviour: literature review’, J. Adv. Nurs., vol. 49, no. 6, pp. 616–623, 2005, doi: 10.1111/j.1365-2648.2004.03331.x.
25. A. Pavli, M. Theodoridou, and H. C. Maltezou, ‘Post-COVID Syndrome: Incidence, Clinical Spectrum, and Challenges for Primary Healthcare Professionals’, Arch. Med. Res., vol. 52, no. 6, pp. 575–581,

Aug. 2021, doi: 10.1016/j.arcmed.2021.03.010.

26. S. A. Reijneveld and K. Stronks, ‘The validity of self-reported use of health care across socioeconomic strata: a comparison of survey and registration data’, *Int. J. Epidemiol.*, vol. 30, no. 6, pp. 1407–1414, Dec. 2001, doi: 10.1093/ije/30.6.1407.
27. M. Voss, S. Stark, L. Alfredsson, E. Vingård, and M. Josephson, ‘Comparisons of self-reported and register data on sickness absence among public employees in Sweden’, *Occup. Environ. Med.*, vol. 65, no. 1, pp. 61–67, Jan. 2008, doi: 10.1136/oem.2006.031427.
28. M. Kroneman, R. Verheij, M. Tacken, and J. van der Zee, ‘Urban–rural health differences: primary care data and self reported data render different results’, *Health Place*, vol. 16, no. 5, pp. 893–902, Sep. 2010, doi: 10.1016/j.healthplace.2010.04.015.
29. I. H. Caspersen, P. Magnus, and L. Trogstad, ‘Excess risk and clusters of symptoms after COVID-19 in a large Norwegian cohort’, Eur. J. Epidemiol., vol. 37, no. 5, pp. 539–548, May 2022, doi: 10.1007/s10654-022-00847-8.
30. R. M. Wong-Chew *et al.*, ‘Symptom cluster analysis of long COVID-19 in patients discharged from the Temporary COVID-19 Hospital in Mexico City’, Ther. Adv. Infect. Dis., vol. 9, p. 20499361211069264, Jan. 2022, doi: 10.1177/20499361211069264.
31. J. W. Blackett, J. Li, D. Jodorkovsky, and D. E. Freedberg, ‘Prevalence and risk factors for gastrointestinal symptoms after recovery from COVID-19’, Neurogastroenterol. Motil., vol. 34, no. 3, p. e14251, 2022, doi: 10.1111/nmo.14251.
32. S. Basharat, Chao, Y, and McGill, S, ‘View of Subtypes of Post–COVID-19 Condition: A Review of the Emerging Evidence’, *Canadian Journal of Health Technologies*, Dec. 2022, Accessed: Dec. 16, 2022. [Online]. Available: https://www.canjhealthtechnol.ca/index.php/cjht/article/view/HC0035/HC0035
33. R. A. Verheij, V. Curcin, B. C. Delaney, and M. M. McGilchrist, ‘Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse’, J. Med. Internet Res., vol. 20, no. 5, p. e185, May 2018, doi: 10.2196/jmir.9134.

Tables

Table 1. Characteristics of GP-EHR cohort

	PCS group – broad definition	PCS group - narrow definition	Non PCS group – broad definition	Non PCS group – narrow definition	Total group individuals with COVID-19 infection
n (% of total COVID-19 group)	3,333 (32.3)	1,533 (14.9)	6,980 (67.7)	8,780 (85.1)	10,313 (100)
Age (mean, SD)	53.0 (18.3)**	53.4 (18.3)**	50.8 (20.1)	51.2 (19.8)	51.5 (19.6)
Children and adolescents, n (%)	71 (2.1)**	22 (1.4)**	380 (5.4)	429 (4.9)	451 (4.4)
Adults, n (%)	2,638 (79.1)**	1,206 (78.7)**	5,326 (76.3)	6,758 (77.0)	7,964 (77.2)
Elderly, n (%)	624 (18.7)**	305 (19.9)**	1,274 (18.3)	1,593 (18.1)	1,898 (18.4)
Male, n (%)	1,127 (33.8%)**	472 (30.8)**	3,111 (44.6)	3,766 (42.9)	4,238 (41.1)
Low education level categories					
Low, n (%)	876 (26.3)	412 (26.9)	1,854 (26.6)	2,318 (26.4)	2,730 (26.5)
Medium, n (%)	579 (17.4)	288 (18.8)	1,179 (16.9)	1,470 (16.7)	1,758 (17.0)
High, n (%)	614 (18.4)	255 (16.6)	1,302 (18.7)	1,661 (18.9)	1,916 (18.6)
Unknown, n (%)	1,264 (37.9)	578 (37.7)	2,645 (37.9)	3,331 (37.9)	3,909 (37.9)
Household income categories					
Low, n (%)	1,298 (38.9)	628 (41.0)*	2,687 (38.5)	3,357 (38.2)	3,985 (38.6)
Medium, n (%)	1,261 (37.8)	567 (37.0)	2,688 (38.5)	3,382 (38.5)	3,949 (38.3)
High, n (%)	630 (18.9)	273 (17.8)	1,305 (18.7)	1,662 (18.9)	1,935 (18.8)
Unknown, n (%)	144 (4.3)	65 (4.2)	300 (4.3)	379 (4.3)	444 (4.3)
Migration background, n	716	353	1377	1740	2.093 (20.3)

(%)	(21.5)*	(23.0)*	(19.7)	(19.8)	
No. of GP consultations 3-12 months after infection (mean, SD)	3.8 (3.5)**	5.5 (4.0)**	0.4 (1.1)	0.8 (1.6)	1.5 (2.7)
No. of different symptoms 3-12 months after infection (mean, SD)	4.6 (4.6)**	6.8 (5.4)**	0.4 (1.2)	0.9 (1.8)	1.8 (3.4)

*p≤0.005 compared to the non-PCS group. **p≤0.001 compared to the non-PCS group; GP = general practitioner. SD = standard deviation

Table 2. Characteristics of the different groups within the Corona Survey Cohort

	PCS	Non – PCS	p-value
n (% of total Corona Survey Cohort included population)	91 (33)	93 (34)	
Age (mean, SD)	53.6 (13.6)	52.3 (12.1)	0.479
Male, n (%)	36 (39.6)	31 (33.3)	0.380
Level of education			≤0.001
Low, n (%)	31 (34.1)	10 (10.8)	
Medium, n (%)	27 (29.7)	33 (35.5)	
High, n (%)	24 (26.4)	48 (51.6)	
Unknown, n (%)	9 (9.9	2 (2.2)	
Migration background. n (%)	8 (8.8)	4 (4.3)	0.422
No. of (self-reported) symptoms			
After 3 months (n=192)	9.2 (4.8)	2.8 (3.4)	≤0.001
After 6 months (n=160)	8.6 (5.4)	2.8 (3.7)	≤0.001
Individuals who are working less or stopped working			
After 3 months (n, % of total n=185)	21 (23.1)	0 (0.0)	≤0.001
After 6 months (n, % of total n=147)	11 (15.5)	1 (1.6)	0.005

Table 3. Characteristics of patients in different symptom clusters

	Group with Community A - psychological-generalized symptoms	Group with Community B – gastrointestinal symptoms	Group with Community C – cardiorespiratory symptoms
n (% of total group included in network analysis. n=1503) with at least two community symptoms	109 (7%)	34 (2%)	18 (1%)
Age (SD)	54.2 (20.0)	51.2 (23.2)	55.3 (21.1)
Female, n (%)	83 (76%)	23 (68%)	13 (72%)
Low education level. n (%)			
High	26 (24%)	n≤10	n≤10
Medium	26 (24%)	14 (41%)	n≤10
Low	17 (16%)	11 (32%)	n≤10
Unknown	40 (36%)	n≤10	n≤10
Migration background. n (%)	26 (24%)	n≤10	n≤10

Figures

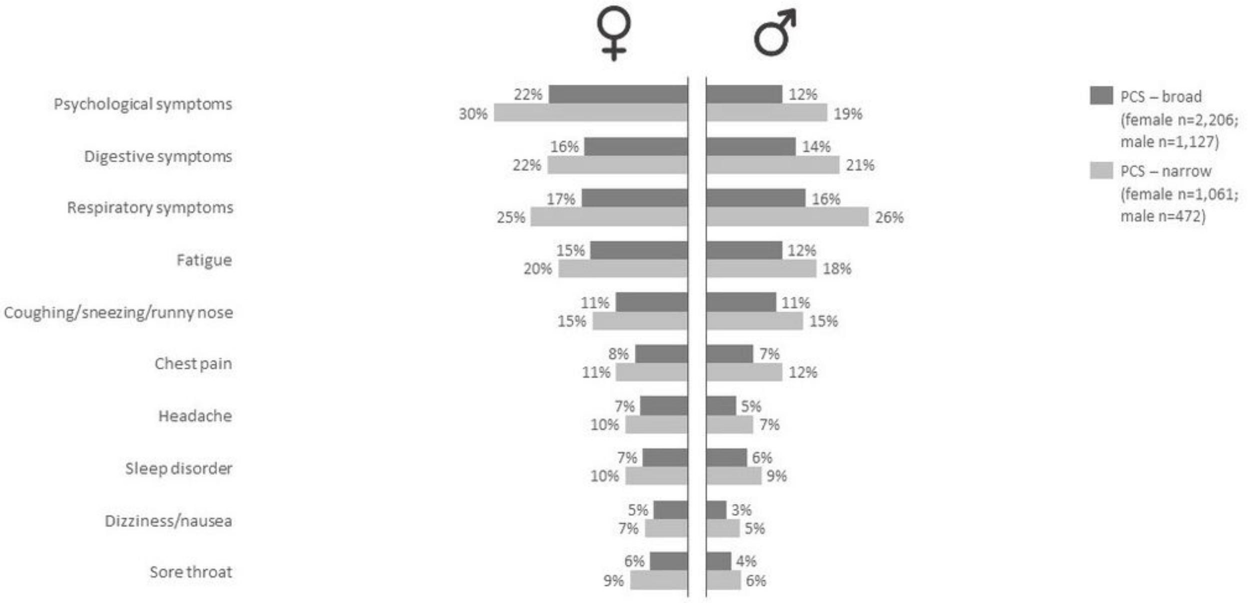

Figure 1

Frequency of symptoms in the GP-EHR cohort stratified by sex.

Barplot showing the frequency of occurrence of category of symptoms in the GP-EHR cohort for the broad (dark grey) and narrow definitions (light grey) stratified for females (left) and males (right).

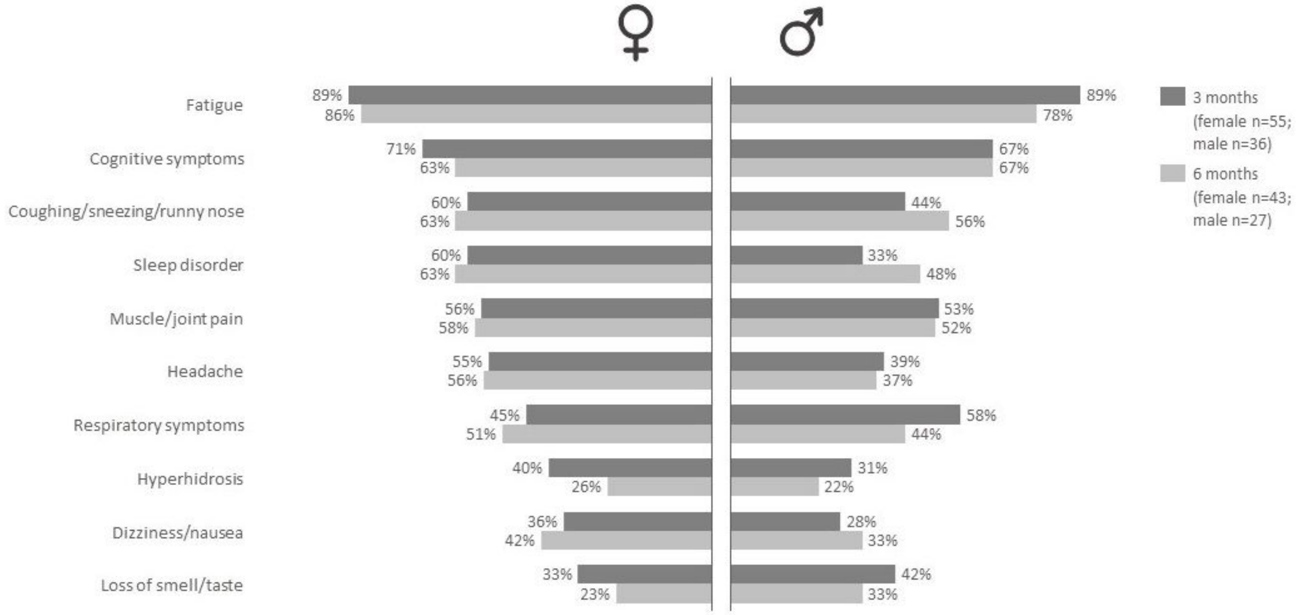

Figure 2

Frequency of symptoms in the Corona Survey Cohort and 3 and 6 months after infections, stratified by sex.

Barplot showing the frequency of self-report symptoms in the Corona Survey Cohort at 3 months (dark grey) and 6 months (light grey) after infection stratified for females (left) and males (right).

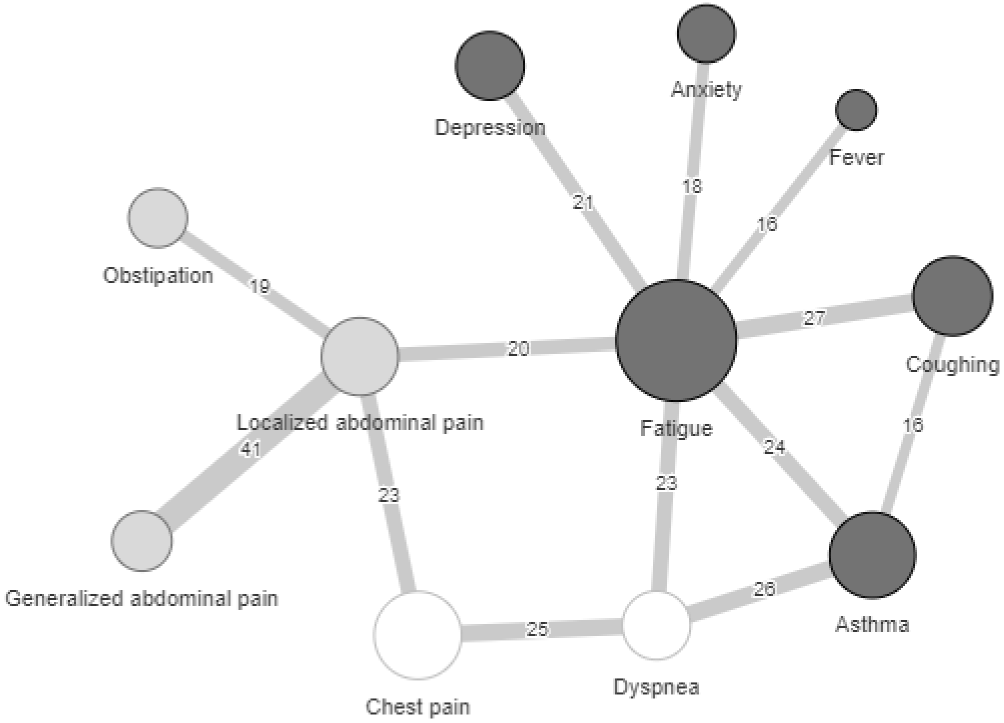

Figure 3

Community detection analyses of symptoms in subgroup of individuals with PCS

Communities of symptoms that co-occur that were detected in a subgroup (n=1503) individuals of the GP-EHR cohort. Three communities were detected which are displayed with different colors: Community A including psychological-generalized symptoms (dark grey), Community B including gastrointestinal symptoms (light grey) and Community C including cardiorespiratory symptoms (white). The size of the circles shows how often symptoms occur in the data with bigger circles occurring more frequent. The numbers indicate how often symptoms co-occur.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx