Statistical Analysis Plan Post-COVID-19 condition on the BES islands

Version 1.0, 10-12-2021

Title of study	Prevalence of and risk factors for Post-COVID-19 Condition (PCC) on the BES islands
Acronym	PCC BES
Trial registration	Not applicable
Study protocol version	Draft
SAP version	1
Statisticians involved	5.1.2e (RIVM) 5.1.2e (RIVM)
Signatures	

Content

1. Introduction

1.1 Background and rationale

Since the start of the COVID-pandemic, Bonaire has seen three waves of increased reports of SARS-CoV-2 infections. Up until October 1st, 2021, roughly 2,077[1] persons living or staying on the island have tested positive for SARS-CoV-2, of whom 57[2] (2.74%) thus far have been admitted to the hospital and 19 (0.91%) have passed due to or with COVID-19. The smaller BES-islands Saba and St Eustatius (Statia) have reported far less cases due to closing their borders early on in the pandemic. A small group of COVID-19 cases continue to report symptoms several weeks following acute COVID-19 infection. This recently recognized medical syndrome, dubbed long COVID, post-COVID-19 condition or long-haul COVID, includes persistence of symptoms like weakness, myalgia, fatigue, shortness of breath and concentration issues following the acute phase of COVID-19; much like other postviral syndromes (Crook et al BMJ 2021, Michelen BMJ Glob Health 2021). We will use the term post-COVID-19 condition for the rest of this SAP. The reported prevalence of post-COVID-19 condition varies widely, depending on the type of COVID-19 patients studied, the criteria used to define post-COVID-19 condition and the moment of measurement (Michelen 2021). Recent estimates of the UK's Office for National Statistics put the prevalence of persistent symptoms beyond 5 weeks at approximately 22.1% and at 9.9% for symptoms lasting beyond 12 weeks.

No research has been carried out focusing on symptomatic prevalence and the severity of post-COVID-19 condition in the Caribbean Region. (Michelen 2021) Secondly, most international post-COVID-19 condition research has been carried out among patients who have been hospitalized or were able to self-report survey data (Crook 2021, Michelen 2021), and are therefore not representative for the general population. Due to their different cultural, religious, and health context, as well as differing health systems, information from the Netherlands or other Western countries on post-COVID-19 condition is not generalizable to the Dutch Caribbean islands. For example, island residents maintain a more passive lifestyle than the European Dutch, are much more often overweight, and hypertension and asthma/COPD are prevalent in the general population (Health Study Caribbean Netherlands 2017); these are important risk factors for severe/fatal COVID-19.

Additionally, COVID-19 aftercare has not been developed fully on the islands, as evidence of what care these patients seek throughout their recovery and rehabilitation is not available. General

[1] Source: Facebook update October 2nd, 2021, Publieke Gezondheid Bonaire.
[2] Source: OSIRIS

practitioners on Bonaire have expressed the need for more insight into the prevalence and symptomatic presentation of post-COVID-19 condition on the island. It is important to determine the source and severity of post-COVID-19 condition symptoms on the BES-islands, and to determine which factors in COVID-patients influence their odds of developing post-COVID-19 condition. Determinants such as severity of the acute disease, lifestyle factors and pre-existing conditions are a few examples.

This research will contribute to the knowledge of post-COVID-19 condition in the Dutch Caribbean and will clarify the health needs of former COVID-19 patients throughout their recovery, which can serve as input for long-term risk management of the public health department and other care providers on the island.

1.2 Research question & Study objectives

This study focuses on answering the following research questions:

- What is the nature and severity of post-COVID-19 condition on the BES-islands?
- What risk factors are associated with developing post-COVID-19 condition among symptomatic COVID-19-patients?

The primary aims of this study are to determine:

1. How many symptomatic persons on the BES-islands experience post-COVID-19 condition;
2. Which persistent symptoms are most frequently reported on the BES-islands;
3. How long people on the BES-islands continue to experience symptoms of COVID-19;
4. To what extent persistent symptoms impact daily functioning and health-related quality of life (HRQoL);
5. To what extent post-COVID-19 condition cases on the BES-islands receive support from their (work) environment;
6. What healthcare needs people with post-COVID-19 condition experience;
7. To what extent severe COVID-19 (defined as hospitalization) is associated with an increased risk of post-COVID-19 condition;
8. Which factors are associated with developing post-COVID-19 condition;
9. Which factors increase the severity of post-COVID-19 condition.

A set of secondary objectives relates to capacity building of the department of Public Health in Bonaire; details are omitted here.

2. Study methods

2.1 Study design

This is a retrospective cohort study on the nature and severity of long term symptoms after COVID-19 infection on the BES islands. There will only be one measurement moment where retrospective data will be collected in the form of telephone interviews (entered through online questionnaires). Apart from the cohort with patients that tested positive for COVID-19, a second cohort of individuals that self-report COVID-19 negative status is be included in the analyses as representation of the general population.

2.2 Subjects

2.2.1 Inclusion criteria for patient cohort

1. Positive COVID-19 test for acute infection (antigen test or PCR)
2. All ages
3. Symptomatic during acute phase
4. Resident of one of the BES-islands

2.2.2 Inclusion criteria for general population cohort

1. Self-reported negative COVID-19 status throughout the epidemic
2. All ages

2.3 Sample size

The sample size calculation was based on the required precision for the estimation of persons suffering from post-COVID-19 condition at 4 weeks or longer after their acute COVID-19 infection. With an alpha of 0.05, power of 80%, expected response rate of 70%, and a conservatively estimated expected post-COVID-19 condition prevalence of 20% at 4 weeks after acute infection, a sample of roughly 600 COVID-19 patients would be needed to estimate the expected 20% prevalence of post-COVID-19 condition with a confidence interval of 6% width (17%-23%). The expectation is that within the patient cohort of 600 COVID-positives, approximately 120 post-COVID-19 condition cases (i.e. 20%) will be identified. This will allow the inclusion of approximately 12 covariates in regression modelling. In addition, a non-patient cohort of 200 self-reported COVID-19 negative persons will be invited in order to be able to compare the distribution of risk factors for and symptoms of post-COVID-19 condition among COVID-19 patients with that of the general population. So in total, 800 persons will be invited to participate in the study.

2.4 Recruitment of participants

Surveillance data will be used to draw a sample of the symptomatic persons infected with SARS-CoV-2 on the BES-islands since April 2020 until October 1st 2021. Symptomatic individuals that have previously tested positive or self-report not having had a COVID-19 infection will be eligible to participate in the study. Data on symptomatic, positively tested individuals' COVID-19 history will be exported from HP-Zone and serve as a sampling frame for the COVID-19 patient cohort. COVID-19 negative controls from the general population will be recruited through referral by cases. These patients and controls will be sent explanation about the study by Whatsapp message and voice note and then contacted by phone and asked to participate in the study.

Figure 1 shows the flow diagram of participant recruitment in the study. Participants are sampled through the following two ways:

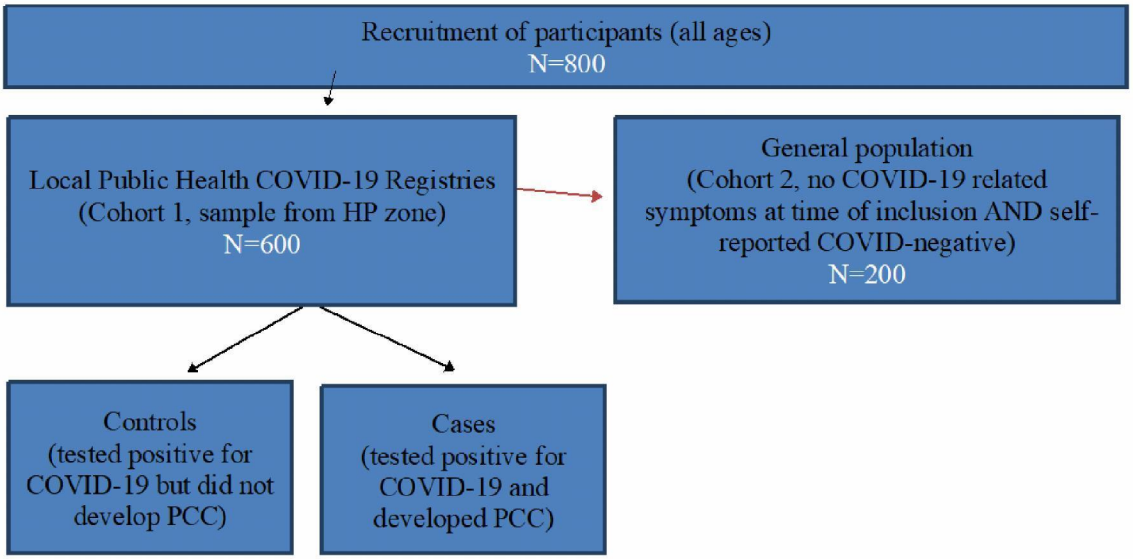
1. *Among people who have tested for COVID-19 in the test street (SARS-CoV-2 patient cohort):*

Staff from the Public Health department in Bonaire will pull a sample of 600 from the HP-Zone export of all symptomatic persons who have tested positive for COVID-19 in the test street since the start of the pandemic. All patients from Saba and Statia (n=xx), all hospitalized patients from all three BES islands (n=57), and a random sample of non-hospitalized Bonaire patients of all ages from the first three waves (n=543) will be drawn from this export, to reach a total of 600 COVID-19 patients.

2. *Through a population sample of adults and children referred by Covid-19 patients (self-reported SARS-CoV-2 negative cohort)*:

At the end of the interview all respondents will be asked to provide contact details of two people who did not get SARS-CoV-2 and who are similar in age and gender and are not from the same household or close contacts. It is expected that this approach will generate a more representative group of the general population as opposed to recruiting individuals who have tested negative at the islands' testing facilities and will lead to a higher participation rate. A random sample of 200 members of this general population sample referred by the Covid-19 cohort participants will be contacted in the same way as the COVID-19 patients.

Figure 1: Recruitment participants post-COVID-19 condition on the BES-islands.



Therefore the study population consists of:

1. COVID-19 patients with a positive test with persisting symptoms lasting for 4 weeks or longer (PCC Cases within Patient Cohort),
2. COVID-19 patients with a positive test and with symptoms lasting for less than 4 weeks (PCC Controls within Patient Cohort),
3. Persons who did not get (tested for) COVID-19 and have no COVID-related symptoms at the time of inclusion (General Population Cohort).

2.5 Informed consent & Ethical approval

Selected participants will be approached through a WhatsApp message and voice note explaining the research, followed by a phone call in which two researchers will ask informed consent from the respondent. Respondents will be explained the purpose of the study and that they can choose to end the interview at any time. A second interviewer will verify whether consent is given and whether the respondent consents to collecting local hospital data, prior to commencing the interview.

As the study focuses on questionnaire data, a non-WMO declaration was obtained, indicating that requesting ethical approval (METC) was not regarded as necessary.

2.6 Data collection

A questionnaire has been developed by the research group, using similar questionnaires by the RIVM, NIVEL, Radboud UMC, GGD Zuid-Holland Zuid, and Maastricht University as guidance. Several stakeholders (GPs, CBS Dutch Caribbean, RIVM colleagues, and public health professionals from all islands) have reviewed the questionnaire, after which it has been piloted by the research team (September 2021) and translated into Spanish, Papiamentu, and English (October 2021). The questionnaire has been developed for three groups: Those who did not have COVID-19 (general population cohort), those who did have COVID-19 but did not develop post-COVID-19 condition (active cases), and those who developed post-COVID-19 condition (PCC patients). The questionnaire focuses on demographics, health status prior to the first COVID-outbreak, vaccination status, symptoms during COVID-infection and during post-COVID-19 condition, quality of life, daily functioning, work/school reintegration and healthcare utilization.

Data will be collected through telephone interviews of approximately 15 minutes for those who did not have COVID-19 and 30 minutes for those who did. It is expected this approach will increase the response rate as opposed to respondents filling in the questionnaire themselves, taking local (health) literacy into consideration. Twenty-four employees of the local public health departments, Red Cross, and/or Tempo organizations, all previously been trained by CBS Dutch Caribbean to carry out phone interviews will conduct these. Participants will be approached between November 15th and 26th, 2021. Data will be directly entered in an existing online data entry system adapted for the purpose of this study (Digital Checklists). This online platform will be managed by one local data manager at the public health department in Bonaire, Ivo Tiemessen. Colleagues from the RIVM will stay in close contact with the data manager to best prepare the dataset in terms of open text answers, formatting, and analysis.

2.7 Definition of outcomes

]

1. The primary outcome variable, **the prevalence of post-COVID-19 condition,** is defined as the presence of at least 1 symptom 4 weeks or longer after receiving a positive test result for SARS-CoV-2, based on the CDC's definition (RQ1, RQ7, RQ8). WHO has recently published their definition of post-COVID-19 condition: "Post COVID-19 condition occurs in individuals with a history of probable or confirmed SARS-CoV-2 infection, usually 3 months from the onset of COVID-19 with symptoms that last for at least 2 months and cannot be explained by an alternative diagnosis. Common symptoms include fatigue, shortness of breath, cognitive

dysfunction but also others which generally have an impact on everyday functioning. Symptoms may be new onset, following initial recovery from an acute COVID-19 episode, or persist from the initial illness. Symptoms may also fluctuate or relapse over time". In order to allow comparison with international studies, we will also calculate the prevalence of post-COVID-19 condition according to this WHO definition, by age and sex categories (RQ1).

2. The second outcome measure is the frequency of individual symptoms (RQ2)
3. The third outcome measure is the duration of PCC symptoms, which is defined in 5 classes: > 1 months and =< 2 months; > 2 months and =<3 months; > 3 months and =< 4 months; > 4 months and =< 6 months; > 6 months
4. The fourth set of outcome measures is **daily functioning and health related quality of life (HRQoL)** (RQ4). ***HRQoL*** will be assessed using the EQ-VAS and the EQ-5D-5L scale[3], considering both the EQ-5D profile distribution and the value set-weighted total EQ-5D score. Because a specific value set for the Dutch Caribbean is not available, the Dutch value set will be used, with a sensitivity analysis using the xxxx value set. ***Daily functioning*** is assessed through a set of questions on reported problems in carrying out daily activities such as self-care, household activities, care of children, sports, hobbies, as well as more specific questions related to impact on work and education.
5. The fifth outcome measure is **support from the (work) environment** (RQ5) , which will be assessed through a) scoring the received support on a scale of 0-4 where 0 represents no sense of support and 4 represents feeling very much supported. Respondents are also able to add to this through an open question. Additionally, respondents are asked if they have received support regarding reintegration to work (yes/no) and if applicable, to specify the type of support received. Additionally, support received from the education environment is asked as well.
6. The sixth outcome measure is **healthcare utilization and need** (RQ6) which is measured in three aspects: 1) as the difference in total number of doctors and/or paramedical practitioners contacted pre-corona and currently (for current PCC patients, former PCC patients and general population controls and 2) total number of visits for medical and paramedical care during the entire PCC episode, additional (non-(para)medical) care use, self-care and location of care (for current and former PCC patients only). For this latter group also the unfulfilled care needs will be assessed.

[3] N. Devlin et al., Methods for Analysing and Reporting EQ-5D Data, https://doi.org/10.1007/978-3-030-47622-9_1

7. The seventh primary outcome measure, the **severity of post-COVID-19 condition**, is defined in two ways, i.e. as duration of complaints and the change in severity score (pre-COVID to during PCC) of four commonly reported symptoms, i.e. shortness of breath, fatigue, brain fog and pain (RQ9).

3. Statistical data analysis

Baseline characteristics of the participants in all groups will be presented using descriptive statistics (mean[standard deviation], median [range] or proportion to assess if there was a balance in the groups regarding distribution of prognostic factors such as age, gender education.

3.1 Primary objectives

3.1.1. How many symptomatic persons on the BES-islands experience post-COVID-19 condition? (RQ1)

Descriptive statistics (prevalence with 95% CI) will identify the proportion of cases (respondents who have tested positive for SARS-COV-2 and fit the definition of post-COVID-19 condition) **overall and by age, sex, BMI and GP practice** and will compare these proportions to those among the general population and among the situation among PCC patients prior to their getting Covid-19. BMI categories will be Underweight < 18, normal 18 to 25, overweight 25 to 30, obesity 30 to 35, morbid obesity >=35 (and age-specific cutoffs for children under 18 years). Differences between groups will be tested with Chi-square and/or Fisher's exact test where appropriate.

3.1.2. Which persistent symptoms are most frequently reported on the BES-islands? (RQ2)

The same analysis as 3.1.1 will identify the frequency of symptoms. These will be described as prevalence with 95% CI and differences between groups tested using Chi-square or Fisher's exact test if appropriate.

Persistent (long-term) symptoms are defined as symptoms persisting for a period of 4 weeks and longer, using a dichotomous outcome (yes/no). The symptoms under question include:

- Shortness of breath;
- Coughing;
- Chest pain;
- Worsened stamina/physical condition;
- Loss of muscle strength;
- Muscle or joint pain;
- Loss of sense of smell;
- Loss of sense of taste;
- Loss of appetite;

- Heart palpitations;
- Brain fog (concentration problems);
- Sleeping problems;
- Fatigue;
- Headache;
- Other complaints

For PCC patients, the severity of individual persistent symptoms (variable set S5_coronaklachten_xxxx) will be described as % in each severity class (mild, moderate, serious, severe) and visualized using absolute and relative stacked bar charts.

3.1.3. How long do people on the BES-islands continue to experience symptoms of COVID-19? (RQ3)

Frequencies of reported duration of persistent symptoms can be calculated from descriptive analysis. Variable '*S5_duur_klachten*' will specify the period in which persons with persistent symptoms continue(d) to experience these after the initial infection.

3.1.4. To what extent do post-COVID-19 condition symptoms impact daily functioning and QoL? (RQ4)

The outcomes of this analysis are

a) health-related quality of life (HRQoL) as measured by EQ-VAS and EQ-5D-5L and

b) daily functioning as measured by a set of questions about the presence and level of problems experienced by PCC patients with regular daily activities such as self-care, household activities, care of children, sports, hobbies, etc. in the past two weeks, as well as more detailed questions about the impact of PCC on a person's work and education.

The EQ-5D-5L and EQ-VAS questionnaire is asked twice: once about the period prior to the coronacrisis and once about the past two weeks prior to the interview. Hence we will be able to compare the EQ-5D-5L profiles and EQ-VAS scores of both current and recovered PCC patients to their own profile prior to the coronacrisis. This will allow analysis of whether the QoL fully rebounds after recovery from PCC. Current PCC patients are those who answer at variable *S5_duur_klachten*: "Ik ben nog niet hersteld van mijn gezondheidsklachten". Recovered PCC patients are those that answer any of the options: "Na 1 maand, maar binnen 2 maanden; Na 2 maanden, maar binnen 3 maanden; Na 3 maanden, maar binnen 4 maanden; Na 4 maanden, maar binnen 6 maanden; Na 6 maanden".

Two approaches will be used to analyse the EQ-5D, i.e. analyzing the change in full profile and in the value set-weighted total score. Because a specific value set for the Dutch Caribbean is not available, and the only value set for other islands in the Caribbean (Trinidad and Tobago) is for the EQ-5D-3L version (not the EQ-5D-5L version we used), we will use the value set for the Netherlands and do a sensitivity analysis using the value set for xxxxx.

The change in EQ-5D-5L profile will be visualized using the Paretian Classification of Health Change (PCHC) as proposed by Devlin et al. (2010). The EQ-VAS results and EQ-5D-5L value set weighted scores will be visualized using histograms and described using parametric (mean, SD) or non-parametric (median, range, IQR) statistics as appropriate. ***The change in EQ-VAS and EQ-5D-5L score*** pre- and post-PCC will be analysed using statistics for paired data (paired t-test or Wilcoxon signed rank test).

These changes in former and current PCC patients will be compared with changes in the general population cohort in order to distinguish the effect of having (had) PCC from the effects of living through lockdowns and other societal measures during the coronacrisis.

Possible determinants of HRQoL loss due to PCC will be explored by comparing HRQoL loss differences (measured by EQ-5D score) among age groups and sociodemographic and clinical characteristics. A change in score of 0.07 will be considered the minimal important difference (Walters 2005).

QALY calculations: The quality-adjusted life year (QALY) is a disease measure indicating the number of years of perfect health lost by a person or group due to mortality or disability (Hollmann 2009). Hence QALY calculations will require data on individual utility loss and the time period for which that loss occurred (i.e., the duration of the Long-COVID) (Hollmann 2009). In our study, the duration of PCC the will be estimated as the mid-point of the answer category of variable *S5_duur_klachten*, e.g. for category "Na 3 maanden, maar binnen 4 maanden" the midpoint will be 3.5 months.

a) Individual loss in terms of QALY (Δ Q) will be calculated for each patient using the following formula (Hollmann 2009);

$$\Delta Q = (\Delta u * d)/365$$

where $\Delta \boldsymbol{u}$ is the utility loss or disutility (HRQoL loss) suffered by the patient, defined as the difference between the EQ-5D score prior to the coronacrisis and the EQ-5D score post PCC; and ***d*** is the duration in days of the persistent symptoms related to PCC.

b) At the population level, the loss in QALYs will be estimated using the mean QALY losses obtained at the individual level

c) QALY loss due to fatal cases will be estimated by imputing the mean population HRQoL by age group at the time of death (Cunillera 2010) and the corresponding life expectancy to the actual number of confirmed deaths caused by Covid-19 in the Dutch Caribbean BES islands.

Impact on daily functioning will be assessed through a set of questions about the presence and level of problems experienced by PCC patients with regular daily activities such as self-care, household activities, care of children, sports, hobbies, etc. in the past two weeks, as well as more detailed questions about the impact of PCC on a person's work and education. The required variables are:

S5_problemen_werk	S5_werk_restklachten
S5_problemen_vrijwilligerswerk	S5_werk_uren
S5_problemen_studie	S5_werkhervatting
S5_problemen_administratie taken	S5_arbeidsongeschikt
S5_problemen_mobiliteit	S5_Verloop_werk
S5_problemen_zelfverzorging	S5_verloop_werk_anders
S5_problemen_huishouden	S5_uitkering

S5_problemen_hobbies	S5_school_restklachten
S5_problemen_sport	S5_school_uren
S5_problemen_sociale contacten	S5_weerschool
S5_problemen_relatie	S5_weerschool_volledig
S5_problemen_verzorging	S5_weerschool_volledig_anders
S5_problemen_mantelzorg	

The variables S5_problemen_x are ordinal variables with answer categories (apart from "niet van toepassing") ranging from "Ik heb/had geen probleem", to "Ik heb/had een beetje problemen", "Ik heb/had matige problemen" and "Ik heb/had ernstige problemen". We will describe these as % of PCC patients in each severity class and visualized using absolute and relative stacked bar charts. The other variables will be analysed using frequency tables or means (SD) / median, range, IQR as appropriate.

Possible determinants of impact on daily functioning due to PCC will be explored by comparing impact differences measured with the "S5_problemen_x" variables among age groups and sociodemographic and clinical characteristics and tested with the Kruskall-Wallis test (or Mann-Whitney-U test in case of only two comparison groups).

3.1.5. To what extent do post-COVID-19 condition patients on the BES-islands receive support from their (work) environment? (RQ5)

The outcome of this analysis is support from the (work) environment related to post-COVID-19 condition (yes/no). The participants are asked if they felt supported by their work environment during the period in which the respondent has (been) experiencing persistent symptoms as well as a follow up question leaving room to clarify why the person did (not) feel they received support.

Support from the work or education environment will be assessed through a) scoring the received support on a scale of 0-4 where 0 represents no sense of support and 4 represents feeling very much supported; and b) if they have received support regarding reintegration to work (yes/no) and if applicable, to specify the type of support received. Additionally**, type of reintegration support received from the education environment** is asked as well and will be analysed in the same way. Variables to be used in this analysis are:

- S5_Steun_werk
- S5_Steun_werk_reden
- S5_janee_hulpverlener
- S5_type hulpverlener
- S5_type hulpverlener_anders
- S5_janee_hulpverleners_school
- S5_type_hulpverlener_school
- S5_type_hulpverlener_school_anders

Differences in support received from the work-environment by age-sex group will be tested using the Kruskall-Wallis test. Reasons for feeling supported (or not) by the work or school environment are open text fields. These will be grouped in themes and presented as line listings.

3.1.6. What healthcare needs do people with post-COVID-19 condition experience? (RQ6)

This question is operationalized as the difference in total number of doctors and/or paramedical practitioners seen pre-corona and currently for current PCC patients, and compared to non-PCC former COVID-19 patients and to general population controls. These comparisons are necessary because the coronacrisis in itself has also impacted on health services availability and peoples' health seeking.

Variables needed for this analysis are:

S2_behandelend_arts_1	S3_behandelend_arts_2
S2_Behandelend_arts_1_anders	S3_Behandelend_arts_2_anders
S2_Paramedisch_1	S3_Paramedisch_2
S2_Paramedisch_1_anders	S3_Paramedisch_2_anders
S2_Behandeling_1	S3_Behandeling_2

The differences in health utilization change between current and recovered PCC patients acute cases, and general population controls will be analyzed using Kruskall-Wallis testing.

The health care use of PCC patients will be further detailed by describing the total number of visits for medical and paramedical care during the entire PCC episode, additional (non-(para)medical) care use, self-care and location of care. Also the unfulfilled care needs will be assessed.

To this end the variables needed are:

S5_afspraken_Acupuncturist	S5_behandeling_restklachten
S5_afspraken_Bedrijfsarts	S5_behandeling_restklachten_anders
S5_afspraken_Cardioloog	S5_zelfzorg_restklachten
S5_afspraken_Cesar therapeut	S5_zelfzorg_restklachten_anders
S5_afspraken_Diëtist	S5_zorgbehoefte_restklachten
S5_afspraken_Ergotherapeut	S5_verzekerdezorg_restklachten
S5_afspraken_Fysiotherapeut	
S5_afspraken_Huisarts	
S5_afspraken_Homeopaat	
S5_afspraken_Internist	
S5_afspraken_KNO-arts	
S5_afspraken_Logopedist	

S5_afspraken_Longarts	
S5_afspraken_Maatschappelijk	
S5_afspraken_Manueel	
S5_afspraken_Psycholoog	
S5_afspraken_Revalidatiearts	
S5_afspraken_Verzekeringsarts	
S5_afspraken_Andere arts	
S5_afspraken_Thuiszorg	

Determinants for high health care utilization by PCC patients will be analyzed using a prediction model with total number of appointments used to construct a dichotomous outcome measure. The study population is all PCC patients.

The number of appointments per healthcare provider will be regarded as being an interval variable, with scores 0 (no appointments had with that particular health care provider), 1 (1-3 appointments); 2 (4-6 appointments); 3 (7-9 appointments) and 4 (10 or more appointments). A total healthcare utilization score during PCC will be calculated by summing the individual provider scores. High healthcare utilization will then be defined as a higher healthcare utilization score than the median use by all PCC patients, i.e. it is a dichotomous variable.

The following variables will be considered as potential predictors:

Required for building the model:	Corresponding variable in questionnaire Voor allemaal geldt: **Filter op PCC_patients**
• Gender	1. S1_geslacht
• Age	2. Age 3. Leeftijdsgroep
• Pre-existing comorbidity	4. S2_Gezondheidsproblemen_1
• Current BMI	5. BMI 6. BMI_category
• Fully vaccinated at time of acute COVID-19 infection (yes/no)	**Maak nieuw variabel:** **FULLY_VAC** = 1 if (difference S6_vaccinatie_datum1 - answer_created_at > 0) AND S6_vaccinatie == "Ja, ik heb COVID-19 gehad, en heb mij daarna 1 keer laten vaccineren" **FULLY_VAC** = 1 if (difference S6_vaccinatie_datum2 -

	answer_created_at > 0) AND S6_vaccinatie == "Ja, ik heb mij al 2 keer laten vaccineren" OR S6_vaccinatie == "Ja, ik heb COVID-19 gehad, en heb mij daarna 2 keer laten vaccineren" **FULLY_VAC ← 0** if (difference S6_vaccinatie_datum1 - answer_created_at =< 0) AND S6_vaccinatie == "Ja, ik ben van plan mij te laten vaccineren" OR S6_vaccinatie == "Ja, en ik ben al opgeroepen, maar ik wil een ander vaccin dan waar ik voor was ingedeeld" OR S6_vaccinatie == Nee OR S6_vaccinatie == "Ik weet (nog) niet of ik mij laat vaccineren"
• Healthcare utilization during acute phase;	Maak nieuw variable **hcuse_acute**: = YES if S4_behandelingcovid != "Ik ben thuis uitgeziekt, en ik heb geen behandeling van mijn huisarts of andere zorgverlener nodig gehad" = NO if S4_behandelingcovid == "Ik ben thuis uitgeziekt, en ik heb geen behandeling van mijn huisarts of andere zorgverlener nodig gehad"
• Hospitalization for acute COVID-19;	PCC_hosp (hier hoef je geen filter te gebruiken)
• Current EQ-VAS;	S3_rapport_2_EQ5D
• Number of persistent symptoms;	
• Severity of persistent fatigue, pain, concentration, shortness of breath;	REMEMBER TO FILTER PCC_PATIENTS S3_LS_vermoeidheid_2 S3_LS_pijn_2 S3_LS_concentratie_2 S3_LS_ademen_2
• Type / duration of treatment received in acute COVID-19 phase	TYPE OF TREATMENT > S4_behandelingcovid DURATION OF TREATMENT (days) > S4_kliniek_duur S4_ic_duur S4_Revalidiecentrum_duur
• GP practice	S1_huisarts (let wel op dat je de recode waarbij de namen van artsen

	zijn weggehaald gebruikt)

After creating/cleaning these variables, we will check for the amount and % of missing data per variable required for this model. The focus will be on missing data in the outcome and predictor variables. We will discuss missing data in other exposure variables as well to check if running a sensitivity analyses would be beneficial. We will use the cutoff value of 5% for evaluating whether an appropriate response to the missing data is required.

In case the missing data for the outcome variable and main predictor variable(s) is >5%, we will look into the pattern of missingness. We will then discuss with the statistician whether multiple imputation or an alternative method is advised to correct for the missing data. In collaboration with the statistician we will run the multiple imputation / alternative as well as develop the prediction model.

Prediction model specification:

To determine the variables to be included in the prediction model to best suit our data, we will employ logistic regression with least absolute shrinkage and selection operator (Lasso) regularization using the full data set as a training set. Lasso regularization encourages models with fewer parameters, thus avoiding overfitting. A sequence of the lasso regularization parameter lambda is considered by the "glmnet" implementation. To determine what value to use for lambda, we'll perform 10-fold cross-validation and identify the lambda value that produces the lowest test mean squared error (MSE). The model's predictive capacity is evaluated with the area under the receiver operating characteristic curve (ROC-AUC) (a.k.a. the c-index).

Performance of the model:

We compare the prediction model to a naïve classification model with the classification probability equal to the frequencies in the data to check whether the data has predictive power. We evaluate the performance of the model on the test set.

***Determining predictors*:**

To assess which variables are important predictors for the prediction model we look at the permutation feature importance (Molnar 2021). We further investigate the marginal effect that important predictors have on the prediction outcome using partial dependence plots (PDP) and Accumulated Local Effects (ALE).

3.1.7. Which factors (including severity of acute COVID-19) predict developing post-COVID-19 condition? (RQ7 + RQ8)

The outcome of this analysis is the dichotomous variable 'post-COVID-19 condition' (yes/no) according to the CDC definition, for which a predictive multivariable logistic regression model with lasso regularization will be built. The model specification and validation approach will be as described under 3.1.5. The study population is restricted to all COVID-19 patients (i.e. only the COVID+ cohort).

Potential predictive factors

Potential predictors are based on previous studies *(add references)*, measured currently (sociodemographics), prior to the start of the coronacrisis (health status related variables, SES, usual health care consumption) or during the acute COVID-19 infection (severity, number of symptoms, health care utilization, Type/duration of treatment period) and include:

- Age in (years)
- Gender (male/female)
- Comorbidity;
- Socioeconomic status;
    - Education level low/moderate/high. Low indicates no education, primary level education, lower vocational and lower secondary education; moderate indicates higher secondary education or undergraduate; high indicates tertiary education, university or postgraduate;
    - Income level pre-COVID period;
- Smoker pre-COVID period (yes/no)
- Drinker pre-COVID period (yes/no)
- BMI-category pre-COVID period;
- Health status pre-COVID period;
    - Coughing, shortness of breath, headache, muscle pain, chest pain, fatigue, eating less, loss of sense of smell, loss of sense of taste, heart palpitations, concentration problems, insomnia, loss of muscle strength, worsened physical condition, no symptoms;
- Healthcare utilization pre-COVID period;
- EQ-VAS pre-COVID
- Healthcare utilization during acute phase;
- Severe COVID, defined as hospitalization;

- Number of symptoms during acute phase;
    - Coughing, shortness of breath, headache, muscle pain, chest pain, fatigue, eating less, loss of sense of smell, loss of sense of taste, heart palpitations, concentration problems, insomnia, loss of muscle strength, worsened physical condition, no symptoms;
- Type / duration of treatment received in acute COVID-19 phase
- Fully vaccinated at time of acute COVID-19 infection (yes/no)

3.1.8. Which factors increase the severity of post-COVID-19 condition? (RQ9)

As far as we know, there is no internationally accepted standard definition of severe PCC. Therefore we will operationalize this question in terms of duration of complaints and the change in severity score of four commonly reported symptoms, i.e. shortness of breath, fatigue, brain fog and pain. Both analyses will use predictive models with the potential predictive factors as specified under 3.1.6 and model building strategies as specified under 3.1.5, except using different link functions (see below).

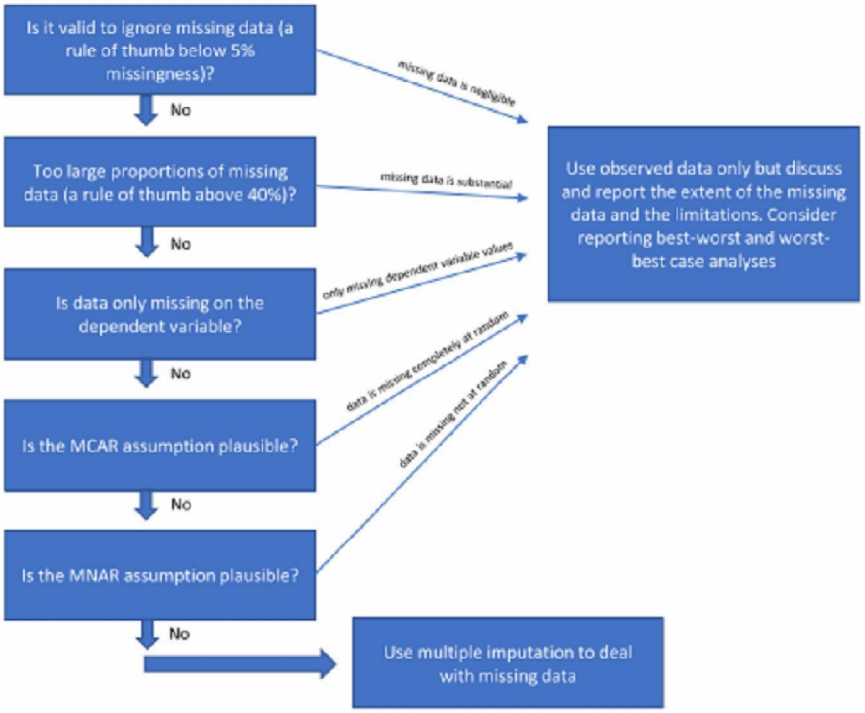
The outcome for the first analysis is an ordinal variable *'S5_duur_klachten'* indicating how long PCC recovery from PCC complaints took, with categories "Na 1 maand, maar binnen 2 maanden; Na 2 maanden, maar binnen 3 maanden; Na 3 maanden, maar binnen 4 maanden; Na 4 maanden, maar binnen 6 maanden; Na 6 maanden". Current PCC patients are those who answer: "Ik ben nog niet hersteld van mijn gezondheidsklachten"- their duration of complaints will be derived from the difference between interview date and date of their positive COVID-19 test and allocated to one of the 5 categories above. The study population for this analysis is all PCC patients.

@choose between dichotomizing the ordinal outcome variable (e.g. < 6 months versus 6+ months) or keeping it ordinal. The former allows logistic regression, the latter would require ordinal logistic regression (if the data satisfy the condition of proportional odds, i.e. the odds (~risk) increases evenly across the categories; this assumption is rarely valid) or generalized ordinal regression / multinomial regression. I am not sure the lasso method is available in R with (generalized) ordinal / multinomial regression, so we might have no choice but to dichotomize the outcome variable.

The outcome for the second analysis is the change in severity of four primary PCC symptoms (shortness of breath, fatigue, brain fog and pain). These were collected about the period of 3 months prior to the COVID-19 complaints and about the week prior to the interview. This change in severity is treated as a continuous outcome, hence we will perform linear regression with lasso regularization for the predictive model. The study population for this analysis is current PCC patients.

3.2. Approach for dealing with missing data

The approach to dealing with missing data for all of the analyses above will depend on the frequency and nature of missingness, and will be guided by the flowchart below.



In case a Complete Case analysis will be done (i.e. only results from records without missings are analyzed), this will be suppl-emented by best-worst case and worst-best case sensitivity analyses. For example for RQ1, to specify the likely true range for the prevalence of post-COVID-19 condition: first a 'best-worst-case' scenario dataset is generated where it is assumed that all COVID-19 patients without outcome information have post-COVID-19 condition symptoms and none of the general population controls do. Then a 'worst-best-case' scenario dataset is generated where it is assumed that none of the COVID-19 patients with missing outcome have post-COVID-19 symptoms and all of the general population controls do.

If the conditions for Complete Case analysis are not met, multiple imputation (MI) will be done. In this approach a large number of imputed datasets are generated using a chosen imputation strategy (e.g. normal model, MICE, predictive mean matching), after which the analysis model is run on all those imputed datasets and results are combined.

@nog toevoegen hoe we beslissen welke MI strategy te gebruiken (zie Harel et al Am J Epi 2017)

To discuss

Nog te bespreken met Bonaire team 5.1.2e

- Akkoord met voorgestelde definitie van severity of PCC?

- Wat doen we met de PEM questions die gesteld zijn onder S3 Huidige gezondheid, want vraagformulering is "over de afgelopen 6 maanden". Dat kan voor PCC patienten bijv een mixed bag zijn van pre-covid, acute covid en PCC
- Een paar andere PEM questions (over effect van mentale en/of lichamelijke inspanning op het verloop van klachten) zijn wel alleen voor PCC patienten gesteld in S5, dus over de situatie tijdens de PCC episode. Er is ook nog een vraag over het verloop van de PCC klachten in het algemeen. Deze toevoegen onder RQ1?
- Wat te doen met de info in de vaccinatievragen (S6) over reden tot niet vaccineren, en de open tekst vraag over effect van vaccinatie op PCC klachten?
- Welke value set te gebruiken voor de sensitivity analysis op de EQ-5D-5L score outcome?
- Welke Multiple Imputation strategy te gebruiken?

References

Nog toevoegen vanuit comment boxes