Phenotyping of acute and persistent COVID-19 features in the outpatient setting: exploratory analysis of an international cross-sectional online survey

Supplementary Material

Health after COVID-19 in Tyrol study team

2021-11-16

Contents

Supplementary Methods

Study design and participants

The bi-national 'Health after COVID-19 in Tyrol' study (ClinicalTrials.gov: NCT04661462) was conducted in the Austrian state of Tyrol and the bordering Italian province of South Tyrol. Between 30th September 2020 and 5th July 2021, COVID-19 convalescents recovering from SARS-CoV-2 infection confirmed by nasal or oral swab PCR or blood antibody test were invited to participate in an anonymized, cross-sectional web-based survey [1] via public media call (local broadcasters: ORF Tirol and RAI Südtirol and newspapers) or contact with a physician. Residency in the study regions and age ≥ 16 (Tyrol) or ≥ 18 years (South Tyrol) were additional study inclusion criteria.

Analysis exclusion criteria in this report were hospitalization because of SARS-CoV-2 infection and the observation time (SARS-CoV-2 test to survey completion) of less than 28 days. Additionally, phenotyping of post-acute sequelae of COVID-19 (PASC) was done with the subsets of the original cohort including the participants with a minimal observation time of 90 days. The scheme of study and analysis inclusion is provided in **Figure 1**.

The study was performed in accordance with the Declaration of Helsinki and the European Data Policy. Digital informed consent was obtained from each participant at the survey start. The study protocols were approved by the institutional review boards of the Medical University of Innsbruck (Tyrol, approval number: 1257/2020) and of the Autonomous Province of South Tyrol/Bolzano (South Tyrol, approval number: 0150701).

Measures

The study questionnaire was developed by a multidisciplinary team (infectious disease specialists, pneumologists, internists, neurologists, psychiatrists, dermatologists, general practitioners, public health and rehabilitation physicians). The survey recorded information on demographics (age, sex, height, weight before infection), socioeconomic status (education, profession, employment status, residence, household size), pre-existing comorbidities (25 items), smoking history, daily medication (quantity, major drug types relevant for SARS-CoV-2 infection course), course of SARS-CoV-2 infection (contact with an infected individual, incubation time, quarantine duration, contact with authorities and physicians), presence and duration of COVID-19 symptoms (44 items), illness perception, symptom relapse as well as psychosocial health and physical constitution during COVID-19 convalescence. Study variables are listed in **Supplementary Table S1**. Baseline demographic, socioeconomic and clinical characteristic of the study populations is presented in **Table 1**. Features of acute COVID-19 course in both study cohorts are presented in **Table 2**. Post-acute characteristic of the study populations is shown in **Table 3**. The German and Italian survey texts as well as the English translation are available as **Supplementary Files**.

Definitions and variable stratification

Respondents were asked to retrospectively assign their COVID-19 symptoms to the following duration classes: absent, present for 1 - 3 days, up to 1 week, up to 2 weeks, up to 4 weeks, up to 3 months, up to 6 months and > 6 months. The surveyed symptoms were classified as (1) acute, when present in the first two weeks after clinical onset, (2) sub-acute when present at 2 - 4 after clinical onset, (3) persistent when present for 4 weeks or longer. Based on the self-reported symptom duration, the individual time intervals till symptom recovery were calculated. Acute COVID-19 was defined as presence of at least one acute symptom, long COVID was defined as presence of at least one persistent symptom for ≥ 28 days, post-acute sequelae of COVID-19 was defined as presence of at least one persistent symptom for ≥ 3 months [2,3].

Symptom relapse after the initial resolution, subjective convalescence and need for rehabilitation were surveyed as single yes/no items each. Illness perception was queried as 'cold-like', 'flu-like', 'gastroenteritis-like' or 'not experienced before/other'. Physical performance loss as compared with the time before SARS-CoV-2 infection was assessed with a 0 - 100 percent scale. Pre-existing comorbidities, depression/anxiety or sleep disorders were surveyed as single yes/no question each. Self-perceived overall mental health (OMH) and quality of life (QoL) impairment were rated with a 4-point Likert scale ('excellent', 'good', 'fair', 'poor', scored: 0, 1, 2, 3). Psychosocial stress was assessed with the PHQ stress module [4,5] without items on weight, sexuality and past traumatic/serious events; the item on worries/dreams was adapted to COVID-19. Participants were stratified by age into young ($\leq$ 30), middle-aged (31 - 65) and elderly ($\geq$ 66 years old). Normal weight, overweight and obesity was defined with 25 and 30 cutoffs of the body mass index (BMI). For modeling tasks, the count of all acute symptoms as well as acute symptoms assigned to the particular phenotypes was stratified by quartiles as presented in **Supplementary Table S2**. For the detailed variable stratification scheme, refer to **Supplementary Table S1**.

Statistical analysis

Data transformation and visualization

Self-reported demographic, biometric, symptom and follow-up data were analyzed and visualized with R version 4.0.5 and *tidyverse* environment [6]. Visualization of the data, modeling and clustering results was done with packages *ggplot2* [7], *cowplot* [8], *ggvenn* and *plotROC* [9]. The entire R analysis pipeline is available at https://github.com/PiotrTymoszuk/health-after-COVID19-analysis-pipeline.

Descriptive statistic and hypothesis testing

For assessing statistical significance of changes in variable frequency between analysis groups or in time, χ^2 test and χ^2 test for trend were applied. To compare differences in medians of numeric variables, Mann-Whitney U test and Kruskal-Wallis test were used, as appropriate for the group number. The R functions used for descriptive statistic and hypothesis testing are available at https://github.com/PiotrTymoszuk/counting-tools. If not indicated otherwise, p values were corrected for multiple testing with Benjamini-Hochberg method [10].

Symptom kinetics modeling

In symptom count modeling, only the participants with the complete set of symptom answers were included. The rates of symptom count reduction in the entire study cohorts in time post clinical onset were modeled with mixed-effect Poisson regression (fixed effect: numeric time after clinical onset, random effect: participant, log-link function) with *glmer()* function from *lme4* package [11]. An analogical modeling approach was utilized to model symptom count kinetic differences in participants suffering from long COVID or PASC with the subsets with complete symptom resolution (fixed effects: numeric time after clinical onset, long COVID/PASC and the time:long COVID/PASC). The exponentiated β_{time} coefficient was interpreted as an estimate symptom resolution rate, the exponentiated $\beta_{interaction}$ was assumed an estimate of symptom reduction rate between the participants with and without long COVID or PASC. Estimate significance was assessed with two-sided T test and degrees of freedom calculated with Satterthwaite method [11,12].

Co-occurrence of symptoms and identification of disease phenotypes

Co-occurrence of acute, persistent (long COVID) and long-term persistent (PASC) symptoms was investigated with pairwise simple matching distances (SMD, function *sm()*, package *nomclust*) [13] and PAM clustering algorithm (partitioning around medoids, function *pam()*, package *cluster*) [14]. Only data from the participants with symptoms at the given time point i.e. acute COVID-19, long COVID and PASC individuals were included in the analysis. The decision on the optimal cluster number was based on the 'bend' of the within-cluster sums-of-square curve (function *fviz_nbclust()*, package *factoextra*, **Supplementary**

Figure S4A), visual analysis of the distance heat maps and results of principal component analysis of the data set (*PCAproj()* function, *pcaPP* package) [15,16]. The R tools for clustering analysis are available at https://github.com/PiotrTymoszuk/clustering-tools-2.

Symptom clusters, further termed 'phenotypes' were defined in the training Tyrol cohort and the symptom - phenotype assignment scheme was applied to the test South Tyrol collective. The general clustering tendency in the training and test cohort was determined with Hopkins statistic (Tyrol: 0.61, 0.66, 0.73, South Tyrol 0.6, 0.73, 0.73 for acute COVID-19, long COVID and PASC, respectively, *get_clust_tendency()* tool, package *factoextra*). The quality and consistency of clustering in the train and test cohorts was assessed by the ratios of between-cluster to total sum of squares (**Supplementary Figure S4B**).

By this means, two phenotypes of acute COVID-19 (NIP: non-specific infection phenotype and MOP: multi-organ phenotype) and three phenotypes of each long COVID and PASC (HAP: hyposmia/anosmia phenotype, FAP: fatigue phenotype, MOP: multi-organ phenotype) were defined. See: **Supplementary Table S4** for the assignment scheme.

Subsets of long COVID and PASC individuals

Association of subjects suffering from long COVID or PASC in respect to the numbers of MOP, FAP and HAP symptoms was explored with pairwise Manhattan distances (function *distance()*, package *philentropy*) [17] and DBSCAN clustering algorithm (function *dbscan()*, package *dbscan*) [18,19]. The minPts argument was set to five based on the $> 2^{data\ dimension}$ rule. The decision on the optimal ϵ parameter value was guided by inspection of the 4 nearest neighbor (4-NN) distance plot (**Supplementary Figure S9A**). The optimal ϵ value was defined as the 4-NN value preceding the steep increase of the 4-NN distance [20].

Definition of participant clusters, further termed 'participant subsets', was done in the training Tyrol cohort. Three participant subsets were identified in long COVID and PASC each: HAP-negative (HAP-), HAP intermediate (HAPi), HAP high (HAP+), termed after on the count of hyposmia/anosmia phenotype symptoms. The subset assignment in the test South Tyrol (STY) cohort was done with k-nearest-neighbor (k = 20 - 50) label propagation algorithm with $dist^{-1}$ kernel-weighted voting [21–23]. The general clustering tendency in the training and test cohort was determined with Hopkins statistic (Tyrol: 0.89 and 0.81, South Tyrol: 0.87 and 0.81 for long COVID and PASC, respectively, *get_clust_tendency()* tool, package *factoextra*). The quality and consistency of clustering in the train and test cohorts was assessed by the ratios of between-cluster to total sum of squares (**Supplementary Figure S9B**) as well as results of principal component analysis of the data set (*PCAproj()* function, *pcaPP* package, **Figure 6A**) [15,16]. The R tools for clustering analysis are available at https://github.com/PiotrTymoszuk/clustering-tools-2.

Uni-variate modeling

Correlation of the candidate factors (**Supplementary Table S7**) with the count of acute COVID-19 symptoms, risk of long COVID and PASC was assessed with a series of univariate generalized linear models (symptom count: Poisson, risk: logistic regression). To account for the sex and age bias as compared with the general population of COVID-19 convalescents, frequency weights were implemented in the modeling procedure based on the publicly available age and sex distributions of the COVID-19 convalescent populations in Tyrol (https://covid19-dashboard.ages.at/dashboard.html, access on 13th July 2021) and Italy (https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_7-luglio-2021.pdf, access on 13th July 2021) for the Tyrol and South Tyrol study cohorts, respectively (**Supplementary Table S8**). To address possible recall bias caused by retrospective surveying of the acute COVID-19 course features, continuous numeric observation time variable (SARS-CoV-2 test to survey compltetion time) was inceluded in every model as a confounder.

Significance of model estimates was determined by Wald Z test and p values were corrected for multiple comparisons with Benjamini-Hochberg method[10]. Model quality and assumptions were checked by visual inspection of the plots of model residuals. The tools used for serial modeling and model quality check are available at https://github.com/PiotrTymoszuk/lm_qc_tools.

LASSO modeling

Identification of the most influential among candidate factors (**Supplementary Table S7**) associated with the count of acute COVID-19 symptoms, risk of long COVID and PASC was accomplished with LASSO (least absolute shrinkage and selection operator) generalized linear modeling (symptom count: Poisson, risk: logistic regression) and *cv.glmnet()* tools from *glmnet* package [24,25]. The models were age- and sex weighted (**Supplementary Table S8**).

LASSO models were constructed in the training Tyrol cohort (**Supplementary Figure S15A** and **S16A**) with the lambda parameter set to 'lambda.1se' resulting in the output models with optimal regularization. The models were subjected to internal (redistribution) and 50-fold cross-validation in the training cohort with the MAE (mean absolute error, symptom count) or MSE (mean squared error, risk) error statistic. Redistribution and cross-validation statistics were extracted with *assess.glmnet()* function and home-developed wrappers (https://github.com/PiotrTymoszuk/lasso_tools). Pseudo R^2 was calculated with the formula $pseudo-R^2 = 1 - deviance/null\ deviance$. Quality of model fit and assumptions were checked by visual inspection of the plots of model residuals.

Symptom count, long COVID and PASC risk predictions by the LASSO models developed in the training Tyrol cohort were externally validated in the South Tyrol test cohort (**Supplementary Figure S15B** and **S16B**). The quality of long COVID and PASC prediction was assessed with receiver-operator characteristic (ROC) using tools proveided by *OptimalCutpoints* and *plotROC* packages [9,26].

Data availability

As this study is still ongoing, the data will be made available on a serious request and made publicly available after the completion. The entire R analysis pipeline is available at https://github.com/PiotrTymoszuk/health-after-COVID19-analysis-pipeline.

Supplementary Tables

Table S1: **Variables queried directly and determined based on survey answers.**
Variable name: full variable name, Variable short name: short variable name used for plot labeling, Unit: variable unit, Description: variable description, Cutpoints: cutoffs used for variable stratification, Levels: variable strata, Variable type: survey module.
The table is available online.

Table S2: **Stratification of the acute symptom count variables for modeling tasks**. #: number, NIP: non-specific infection phenotype, MOP: multi-organ phenotype.

Cohort	Quartile	# acute symptoms	# acute NIP symptoms	# acute MOP symptoms
North Tyrol	Q1	(0, 9]	(0, 7]	(0, 1]
	Q2	(9, 13]	(7, 10]	(1, 3]
	Q3	(13, 18]	(10, 12]	(3, 6]
	Q4	(18, 42]	(12, 16]	(6, 26]
South Tyrol	Q1	(0, 7]	(0, 6]	(0, 1]
	Q2	(7, 13]	(6, 9]	(1, 3]
	Q3	(13, 18]	(9, 12]	(3, 7]
	Q4	(18, 39]	(12, 16]	(7, 23]

Table S3: **Whole-cohort symptom prevalence and time changes of COVID-19 symptoms**
Symptom frequency at the given time point (first two, two to four and four weeks or longer after symptom onset) was determined as a percent of the entire cohort. Statistical significance of the time change in symptom frequency was assessed by χ^2 test for trend and p values corrected for multiple comparisons with Benjamini-Hochberg method.
Time point: time interval after symptom onset, N: number of cohort members with the symptom.
The table is available online.

Table S4: **Assignment of acute and persistent COVID-19 symptoms to the phenotypes.** imp.: impaired, dim.: diminished.

Time point	Phenotype	Symptoms
acute COVID-19	Non-specific Infection Phenotype (NIP)	fever, sore throat, running nose, fatigue, dry cough, tachypnea, chest pain, joint pain, muscle pain, dim. appetite, dizziness, headache, hypo/anosmia, hypo/ageusia, tiredness at day, imp. concentration
	Multi-Organ Phenotype (MOP)	shivering, wet cough, dyspnea, tachycardia, palpitations, bone pain, abdominal pain, nausea, vomiting, diarrhea, confusion, tingling feet, tingling hands, burning feet, burning hands, numb feet, numb hands, imp. walk, imp. fine motor skills, sleeplessness, forgetfulness, epilepsy, swelling, blue fingers/toes, urticaria, blistering rash, blue marmorated skin, red eyes
long COVID	Hyposmia/Anosmia Phenotype (HAP)	hypo/anosmia, hypo/ageusia
	Fatigue Phenotype (FAP)	fatigue, tiredness at day, imp. concentration, forgetfulness
	Multi-Organ Phenotype (MOP)	fever, shivering, sore throat, running nose, dry cough, wet cough, tachypnea, dyspnea, chest pain, tachycardia, palpitations, joint pain, bone pain, muscle pain, abdominal pain, nausea, vomiting, dim. appetite, diarrhea, dizziness, headache, confusion, tingling feet, tingling hands, burning feet, burning hands, numb feet, numb hands, imp. walk, imp. fine motor skills, sleeplessness, epilepsy, swelling, blue fingers/toes, urticaria, blistering rash, blue marmorated skin, red eyes
PASC	Hyposmia/Anosmia Phenotype (HAP)	hypo/anosmia, hypo/ageusia
	Fatigue Phenotype (FAP)	fatigue, tachypnea, tiredness at day, imp. concentration, forgetfulness
	Multi-Organ Phenotype (MOP)	fever, shivering, sore throat, running nose, dry cough, wet cough, dyspnea, chest pain, tachycardia, palpitations, joint pain, bone pain, muscle pain, abdominal pain, nausea, vomiting, dim. appetite, diarrhea, dizziness, headache, confusion, tingling feet, tingling hands, burning feet, burning hands, numb feet, numb hands, imp. walk, imp. fine motor skills, sleeplessness, epilepsy, swelling, blue fingers/toes, urticaria, blistering rash, blue marmorated skin, red eyes

Table S5: **Demographic and clinical characteristic of the long COVID and PASC participant subsets.**
HAP neg: hypo/anosmia-negative, HAP int: hypo/anosmia intermediate, HAP pos: hypo-anosmia high phenotype subset, Raw p: unadjusted p value obtained with χ^2 test, Adjusted p: p value adjusted for multiple testing with Benjamini-Hochberg method.
The table is available online.

—

—

—

Table S6: **Physical, quality of life and mental health inpairment scoring in the long COVID and PASC participant subsets.**
HAP neg: hypo/anosmia-negative, HAP int: hypo/anosmia intermediate, HAP pos: hypo-anosmia high phenotype subset, Raw p: unadjusted p value obtained with Kruskal-Wallis test, Adjusted p: p value adjusted for multiple testing with Benjamini-Hochberg method.
The table is available online.

—

—

—

Table S7: **Candidate variables in modeling tasks.**.

Response	Method	Co-variates
# acute symptoms	GLM Poisson	Sex, Age, BMI before CoV, Autoimmunity, Anemia, Hypertension, Pre-CoV depr/anxiety, Diabetes, Surg. Before CoV, Freq. resp. inf., CVD, Allergy, Malignancy, Paresthesia, GI disease, Lung dis., Freq. bact. Inf, Sleep apnea, Bruxism, CKD, Pre-CoV sleep disord., Embolism, # comorb., Daily medic., Smoking
long COVID	logistic regression	Sex, Age, BMI before CoV, Autoimmunity, Anemia, Hypertension, Pre-CoV depr/anxiety, Diabetes, Surg. Before CoV, Freq. resp. inf., CVD, Allergy, Malignancy, Paresthesia, GI disease, Lung dis., Freq. bact. Inf, Sleep apnea, Bruxism, CKD, Pre-CoV sleep disord., Embolism, # comorb., Daily medic., Smoking, CoV no therapy, CoV anti-pyretic, CoV antibiotic, Acute fever, Acute shivering, Acute sore throat, Acute running nose, Acute fatigue, Acute dry cough, Acute wet cough, Acute tachypnea, Acute dyspnea, Acute chest pain, Acute tachycardia, Acute palpitations, Acute joint pain, Acute bone pain, Acute muscle pain, Acute abd. pain, Acute nausea, Acute dim. appetite, Acute diarrhea, Acute dizziness, Acute headache, Acute hyposmia/anosmia, Acute hypogeusia/ageusia, Acute confusion, Acute tingling feet, Acute tingling hands, Acute imp. walk, Acute sleeplessness, Acute tiredness at day, Acute imp. concentration, Acute forgetfulness, Acute red eyes, # acute symptoms, # acute NIP symptoms, # acute MOP symptoms
PASC	logistic regression	Sex, Age, BMI before CoV, Autoimmunity, Anemia, Hypertension, Pre-CoV depr/anxiety, Diabetes, Surg. Before CoV, Freq. resp. inf., CVD, Allergy, Malignancy, Paresthesia, GI disease, Lung dis., Freq. bact. Inf, Sleep apnea, Bruxism, CKD, Pre-CoV sleep disord., Embolism, # comorb., Daily medic., Smoking, CoV no therapy, CoV anti-pyretic, CoV antibiotic, Acute fever, Acute shivering, Acute sore throat, Acute running nose, Acute fatigue, Acute dry cough, Acute wet cough, Acute tachypnea, Acute dyspnea, Acute chest pain, Acute tachycardia, Acute palpitations, Acute joint pain, Acute bone pain, Acute muscle pain, Acute abd. pain, Acute nausea, Acute dim. appetite, Acute diarrhea, Acute dizziness, Acute headache, Acute hyposmia/anosmia, Acute hypogeusia/ageusia, Acute confusion, Acute tingling feet, Acute tingling hands, Acute imp. walk, Acute sleeplessness, Acute tiredness at day, Acute imp. concentration, Acute forgetfulness, Acute red eyes, # acute symptoms, # acute NIP symptoms, # acute MOP symptoms

Table S8: **Weights applied to age and strata in modeling tasks.**.
Convalescents: number of convalescents in the given strata in the Tyrol or Italy population.

Cohort	Age strata	Sex	Convalescents	Freq. weight
North Tyrol	<5	male	492	0.0078720
	<5	female	402	0.0064320
	5-14	male	2750	0.0440000
	5-14	female	2574	0.0411840
	15-24	male	4846	0.0775360
	15-24	female	4478	0.0716480
	25-34	male	5241	0.0838560
	25-34	female	4960	0.0793600
	35-44	male	4411	0.0705760
	35-44	female	4711	0.0753760
	45-54	male	5209	0.0833440
	45-54	female	5658	0.0905280
	55-64	male	4467	0.0714720
	55-64	female	4053	0.0648480
	65-74	male	1875	0.0300000
	65-74	female	2000	0.0320000
	75-84	male	1294	0.0207040
	75-84	female	1597	0.0255520
	>84	male	474	0.0075840
	>84	female	1008	0.0161280
South Tyrol	0-9	male	120929	0.0293421
	10-19	male	214102	0.0519495
	20-29	male	256298	0.0621879
	30-39	male	258715	0.0627743
	40-49	male	324116	0.0786431
	50-59	male	358095	0.0868878
	60-69	male	233361	0.0566225
	70-79	male	151073	0.0366562
	80-89	male	76247	0.0185005
	Over 90	male	12787	0.0031026
	0-9	female	112727	0.0273520
	10-19	female	196809	0.0477535
	20-29	female	248432	0.0602793
	30-39	female	271502	0.0658769
	40-49	female	356740	0.0865590
	50-59	female	373925	0.0907287
	60-69	female	221200	0.0536717
	70-79	female	157086	0.0381152
	80-89	female	126781	0.0307620
	Over 90	female	50426	0.0122353

Table S9: **Correlation of candidate factors with the count of acute COVID-19 symptoms, long COVID and PASC risk in univariate modeling.**
Exp. estimate: exponentiated regression estimate with 95% confidence interval, $exp\beta$ for the symptom count and odds ratio (OR) for the risk modeling, Significance: p value adjusted for multiple testing with Benjamini-Hochberg method.
The table is available online.

Supplementary Figures

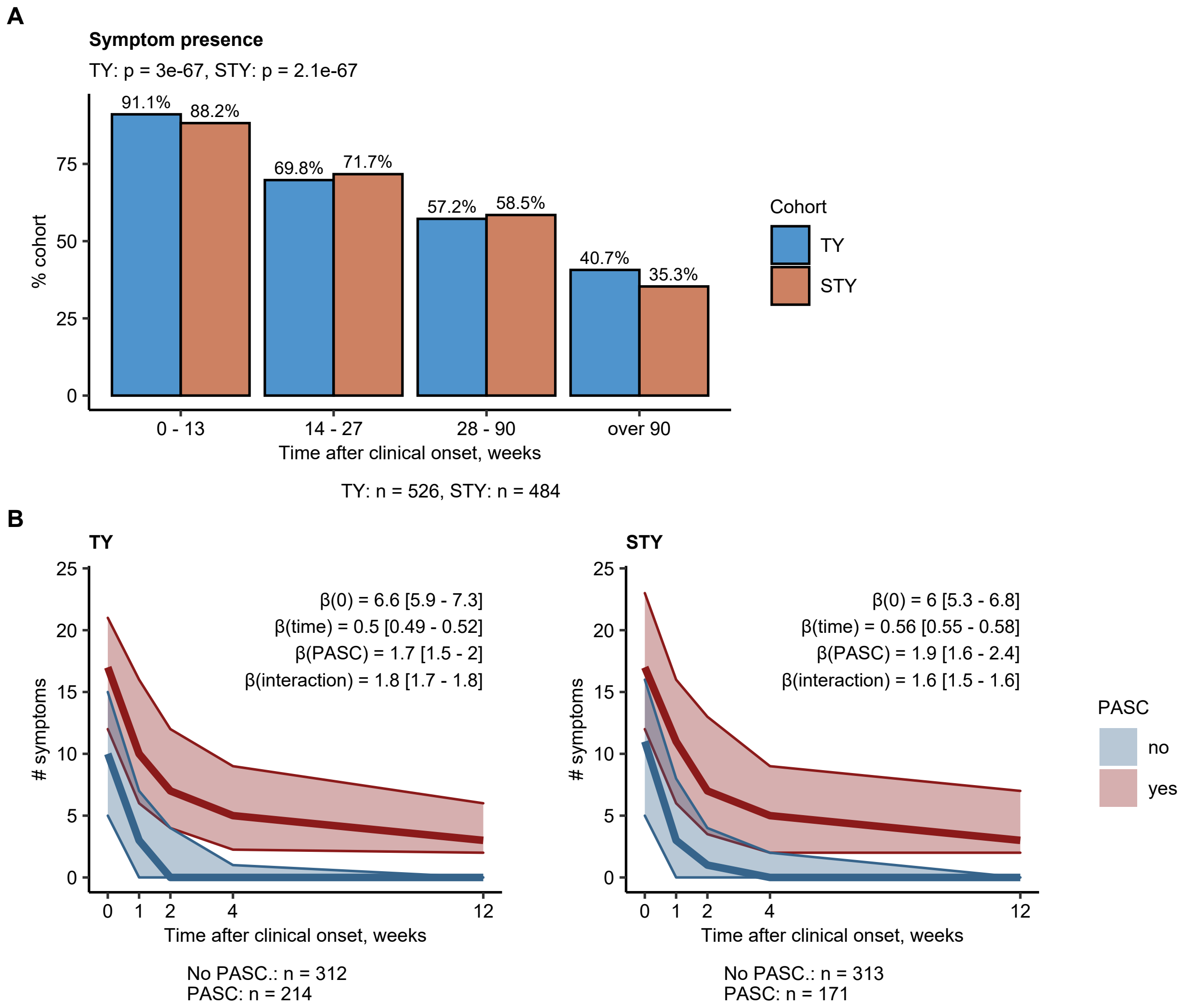


Figure S1: Kinetic of symptom resolution in PASC.

Supplementary Figure S1. Kinetic of symptom resolution in PASC.

Frequency of symptomatic individuals and symptom resolution kinetics were investigated in the subsets of the Tyrol (TY) and South Tyrol (STY) cohorts with the minimal observation time (SARS-CoV-2 test - survey) of 90 days.

(A) Percents of symptomatic participants in time. Statistical significance was determined by χ^2 test for trend. P values are shown in the plot caption.

(B, C) Symptom number trajectories in the entire study cohorts (**B**) and in the subsets with or without PASC. Thin gray lines: individual symptom number trajectories, thick color line: median symptom count, color ribbon: IQR. Statistical significance was determined by mixed-effect Poisson modeling. Model estimates (β) with 95% CI and p values are indicated in the plot.

Numbers of complete cases are indicted under the plots.

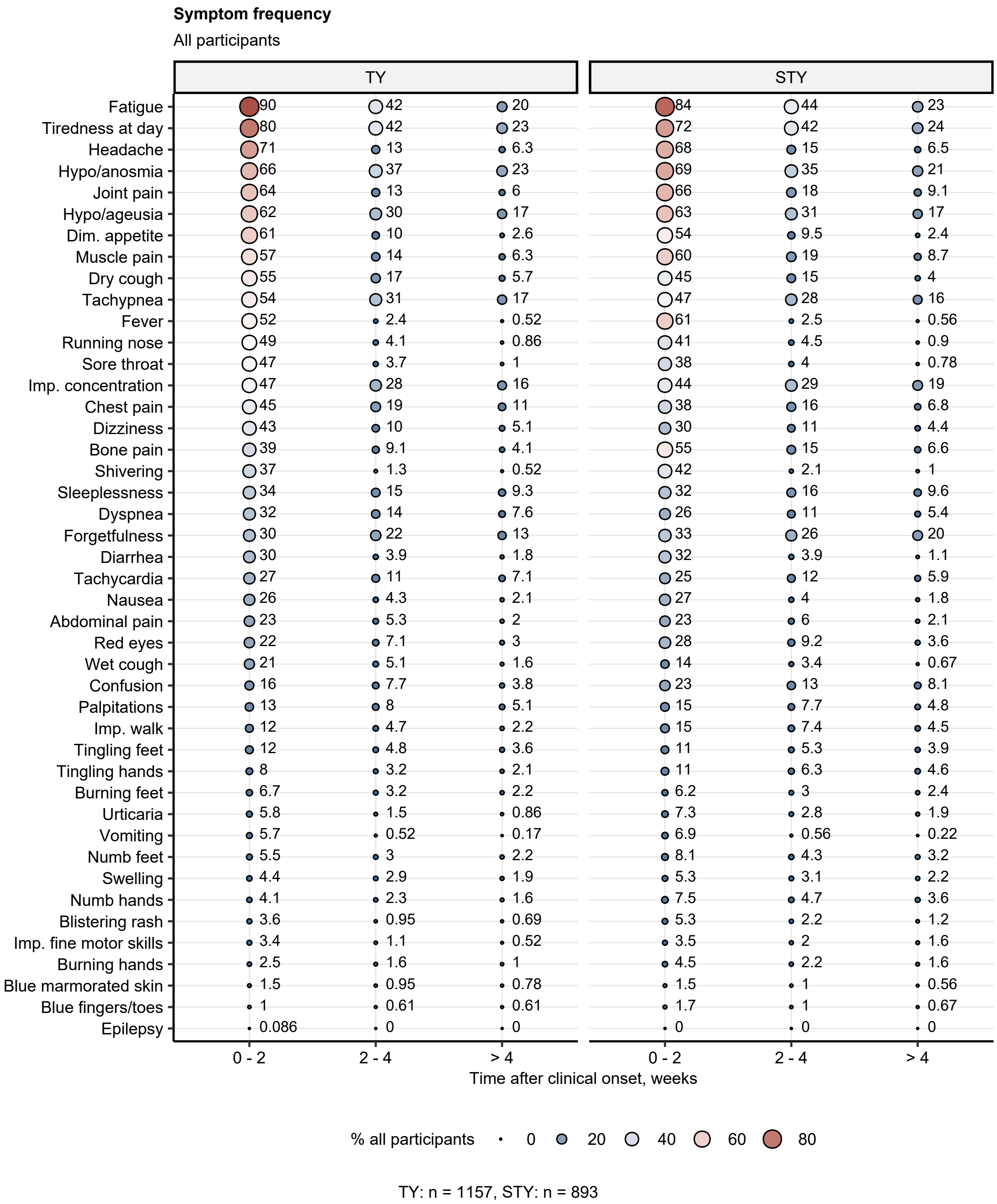


Figure S2: Symptom frequency in acute and sub-acute COVID-19 and long COVID in the entire study cohorts.

Supplementary Figure S2. Symptom frequency in acute and sub-acute COVID-19, long COVID and PASC in the entire study cohorts.

Symptom frequencies were expressed as percentages of the respective study cohort. Point size and color represents the percentage. Numbers of complete observations are indicated below the plot.

tired. day: tiredness at day, imp.: impaired, conc.: concentration, abd. pain: abdominal pain, dim.: diminished, f.m.s: fine motor skills, bl.: blue, marm. skin: marmorated skin, TY: Tyrol, STY: South Tyrol cohort.

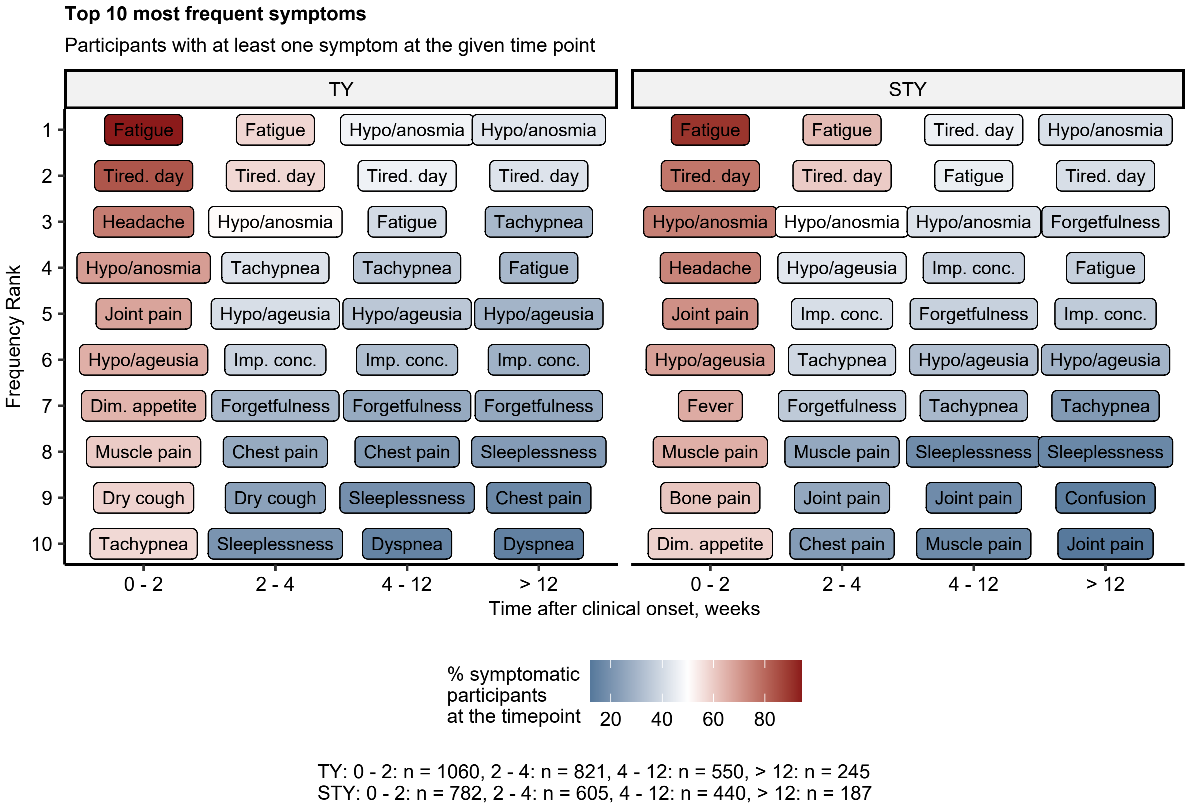


Figure S3: Ten most frequent symptoms of acute, sub-acute COVID-19, long COVID and PASC.

Supplementary Figure S3. Ten most frequent symptoms of acute, sub-acute COVID-19, long COVID and PASC.

Symptom frequencies were expressed as percentages of the individuals with symptoms at the indicated time points after clinical onset. Ten most frequent symptoms at the indicated time points are presented. Numbers of complete observations are indicated below the plot.

tired. day: tiredness at day, imp.: impaired, conc.: concentration, dim.: diminished, TY: Tyrol, STY: South Tyrol cohort.

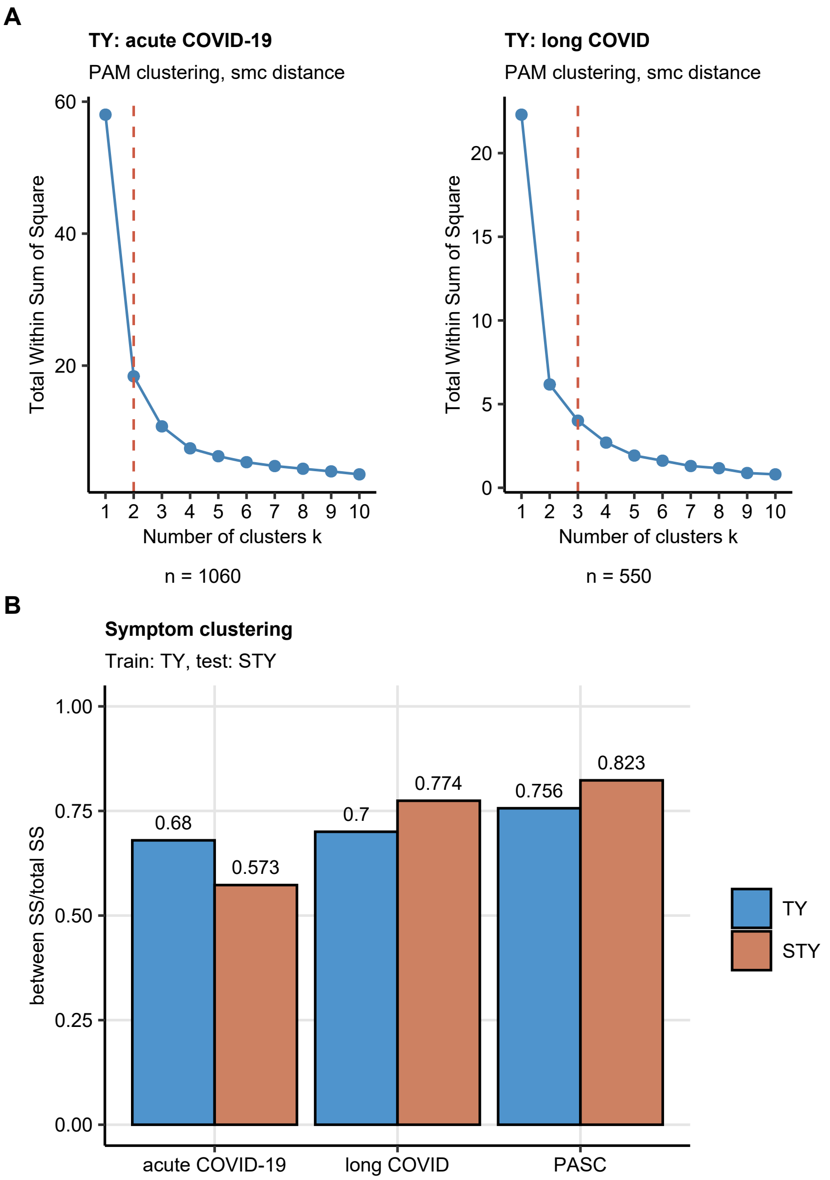


Figure S4: Determination of the optimal cluster number and clustering variance in association analysis of acute COVID-19, long COVID and PASC symptoms.

Supplementary Figure S4. Determination of the optimal cluster number and clustering variance in association analysis of acute COVID-19, long COVID and PASC symptoms.

Association of acute COVID-19, long COVID and PASC symptoms in the training Tyrol (TY) cohort was investigated by simple matching distance (SMD) and PAM (partitioning around medoids) algorithm. The phenotype assignment scheme was applied to the test South Tyrol data set.

(A) Plots of total within-cluster sum of squares as a function of cluster number used to guide the decision on the optimal cluster count by the 'curve elbow' method. Dashed vertical lines represent the chosen numbers of persistent symptom clusters.

(B) Ratios of between-cluster to total sum of squares (SS).

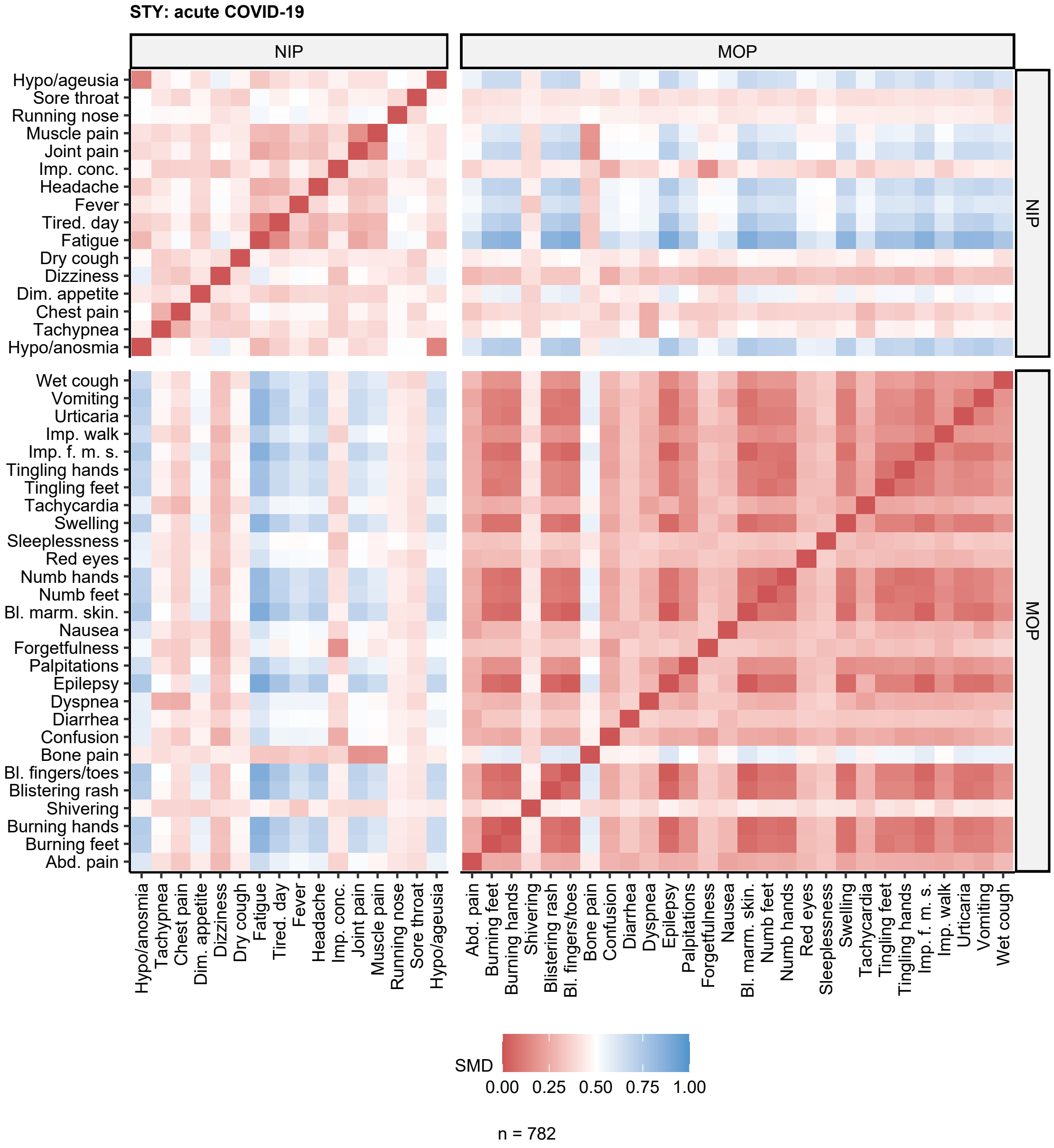


Figure S5: Clustering of acute COVID-19 symptoms in the test South Tyrol cohort.

Supplementary Figure S5. Clustering of acute COVID-19 symptoms in the test South Tyrol cohort.

Clusters (phenotypes) of acute COVID-19 symptoms, the non-specific infection (NIP) and multi-organ phenotype (MOP), were defined in the training Tyrol (TY) cohort by simple matching distance (SMD) and PAM (partitioning around medoids) algorithm. The phenotype assignment scheme was applied to the test South Tyrol (STY) data set. SMD values for symptom pairs in the STY cohort are presented as a heat map. The number of complete observations is indicated under the plot.

tired. day: tiredness at day, imp.: impaired, conc.: concentration, abd. pain: abdominal pain, dim.:

diminished, f.m.s: fine motor skills, bl.: blue, marm. skin: marmorated skin.

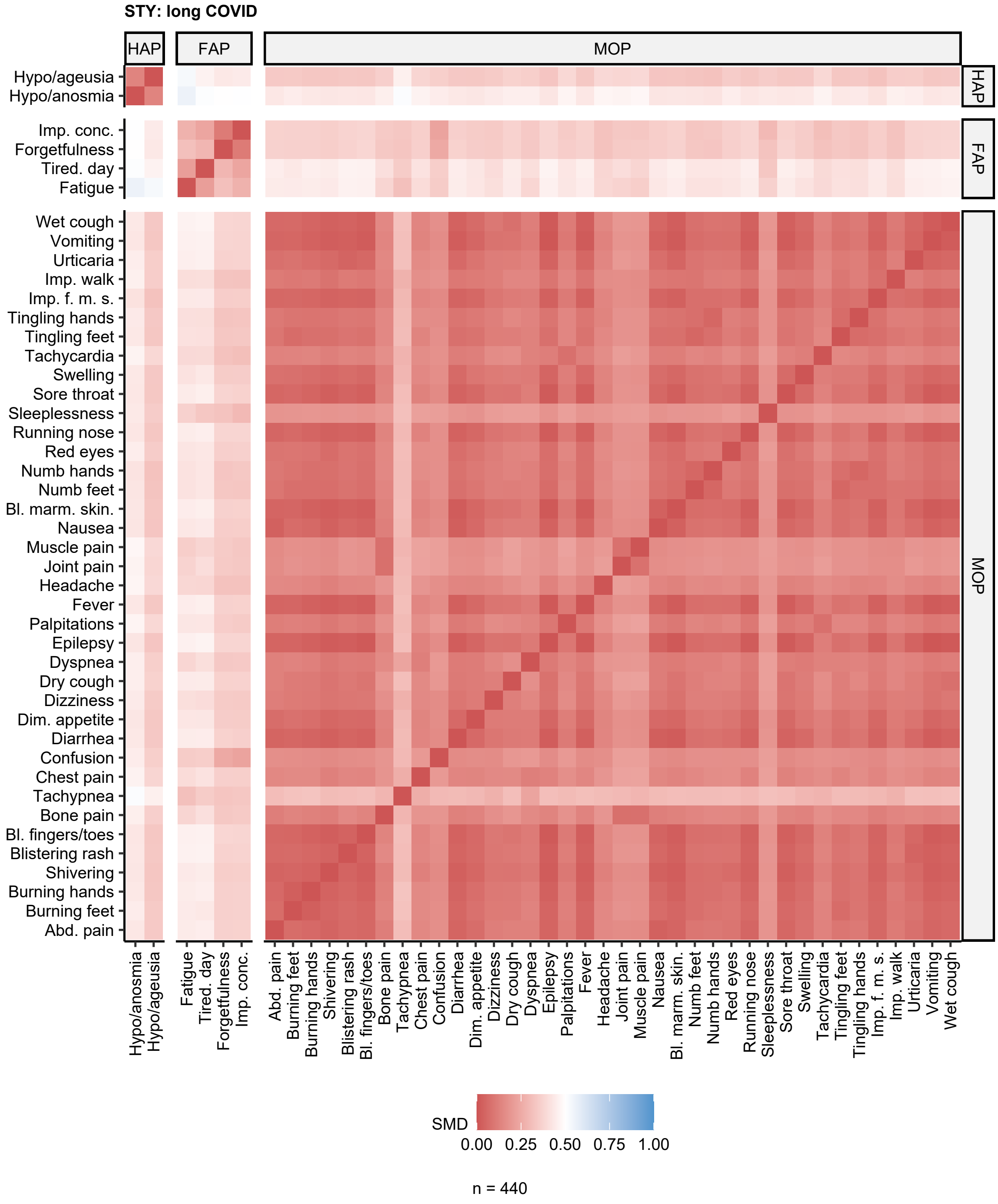


Figure S6: Clustering of long COVID symptoms in the test South Tyrol cohort.

Supplementary Figure S6. Clustering of long COVID symptoms in the test South Tyrol cohort.

Clusters (phenotypes) of long COVID symptoms, the hypo/anosmia (HAP), fatigue (FAP) and multi-organ phenotype (MOP), were defined in the training Tyrol (TY) cohort with simple matching distance (SMD) and PAM algorithm. The phenotype assignment scheme was applied to the test South Tyrol (STY) data set. SMD values for symptom pairs in the STY cohort are presented as a heat map. The number of complete

observations is indicated under the plot.

tired. day: tiredness at day, imp.: impaired, conc.: concentration, abd. pain: abdominal pain, dim.: diminished, f.m.s: fine motor skills, bl.: blue, marm. skin: marmorated skin.

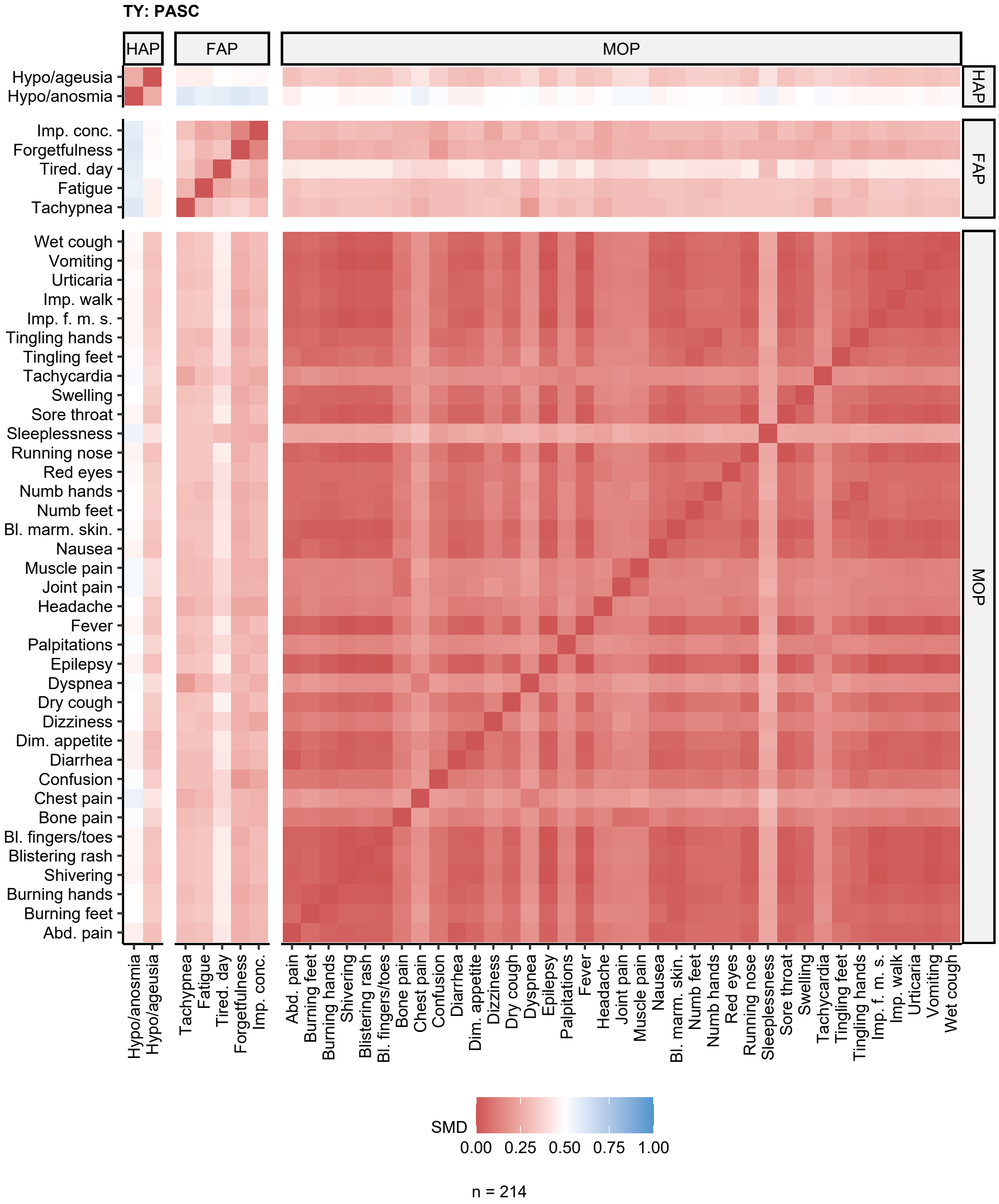


Figure S7: Clustering of PASC symptoms in the training Tyrol cohort.

Supplementary Figure S7. Clustering of PASC symptoms in the training Tyrol cohort.

Clusters (phenotypes) of PASC symptoms, the hypo/anosmia (HAP), fatigue (FAP) and multi-organ phenotype (MOP), were defined in the training Tyrol (TY) cohort with simple matching distance (SMD) and PAM algorithm. The phenotype assignment scheme was applied to the test South Tyrol (STY) data set. SMD values for symptom pairs in the TY cohort are presented as a heat map. The number of complete

observations is indicated under the plot.

tired. day: tiredness at day, imp.: impaired, conc.: concentration, abd. pain: abdominal pain, dim.: diminished, f.m.s: fine motor skills, bl.: blue, marm. skin: marmorated skin.

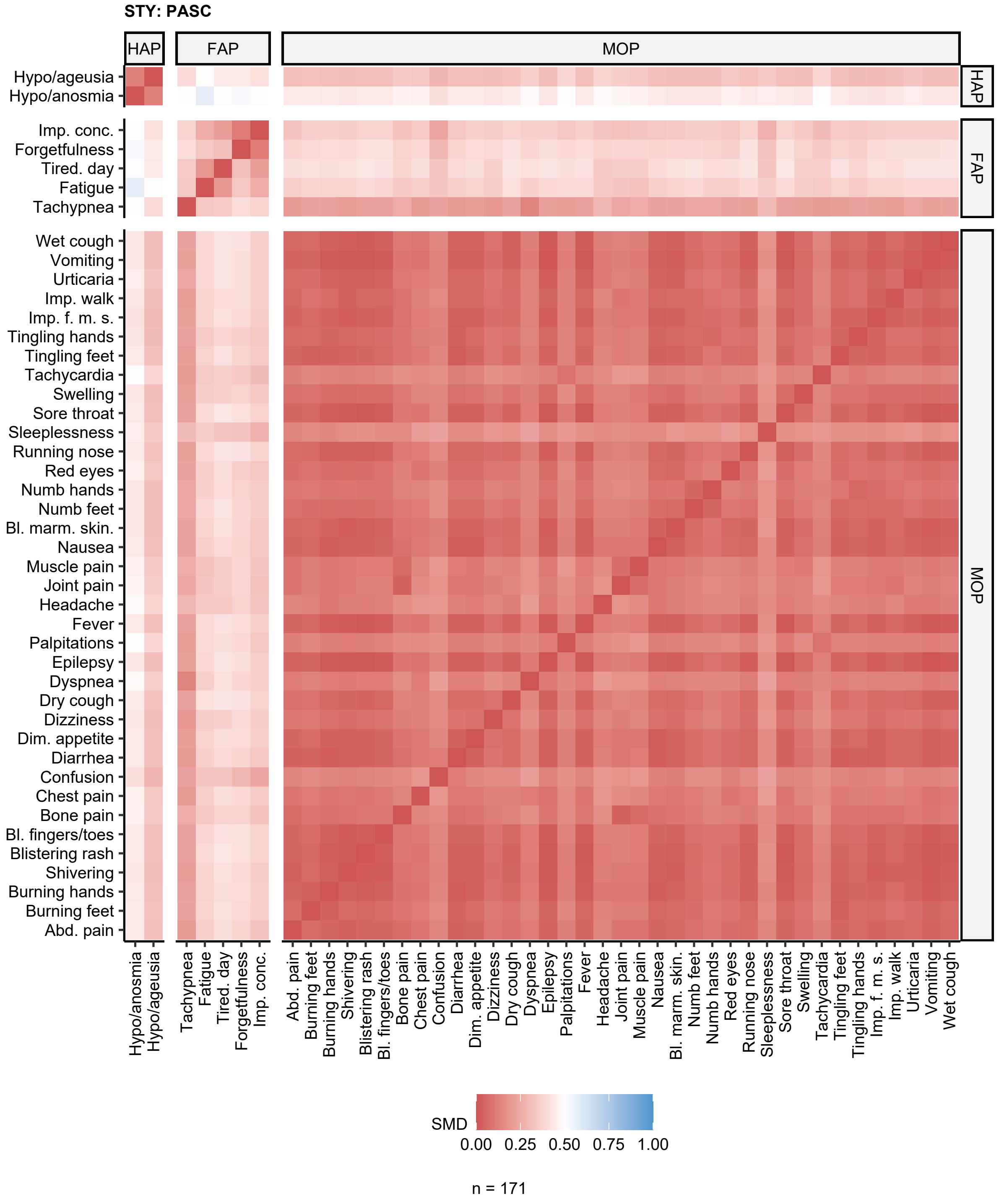


Figure S8: Clustering of PASC symptoms in the test South Tyrol cohort.

Supplementary Figure S8. Clustering of PASC symptoms in the test South Tyrol cohort.

Clusters (phenotypes) of PASC symptoms, the hypo/anosmia (HAP), fatigue (FAP) and multi-organ phenotype (MOP), were defined in the training Tyrol (TY) cohort with simple matching distance (SMD) and PAM algorithm. The phenotype assignment scheme was applied to the test South Tyrol (STY) data set. SMD values for symptom pairs in the STY cohort are presented as a heat map. The number of complete

observations is indicated under the plot.

tired. day: tiredness at day, imp.: impaired, conc.: concentration, abd. pain: abdominal pain, dim.: diminished, f.m.s: fine motor skills, bl.: blue, marm. skin: marmorated skin.

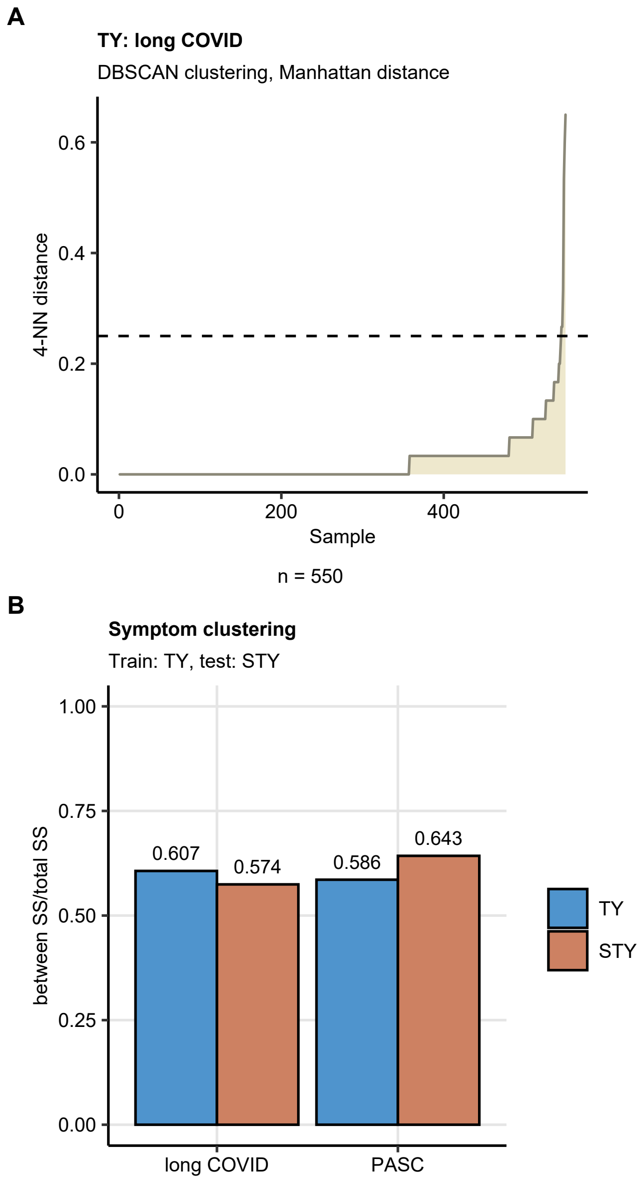


Figure S9: Determination of the optimal ϵ parameter value and clustering variance in association analysis of long COVID and PASC individuals.

Supplementary Figure S9. Determination of the optimal ϵ parameter value and clustering variance in association analysis of long COVID and PASC individuals.

Subsets of long COVID and PASC individuals were defined in the training Tyrol (TY) cohort with Manhattan distance and DBSCAN clustering according to the counts of hypo/anosmia (HAP), fatigue (FAP) and multi-organ phenotype (MOP) symptoms. The subset assignment in the test South Tyrol (STY) cohort was done with k-nearest-neighbor label propagation algorithm. The analysis was conducted in the subsets of the study cohorts with the minimal observation time (SARS-CoV-2 test - survey) of 90 days.

(A) Plots of the sorted 4-nearest neighbor (4-NN) Manhattan distances used to guide the decision on the optimal value of the ϵ. The optimal ϵ value was defined as the 4-NN value preceding the steep increase of the 4-NN distance.

(B) Ratios of between-cluster to total sum of squares (SS).

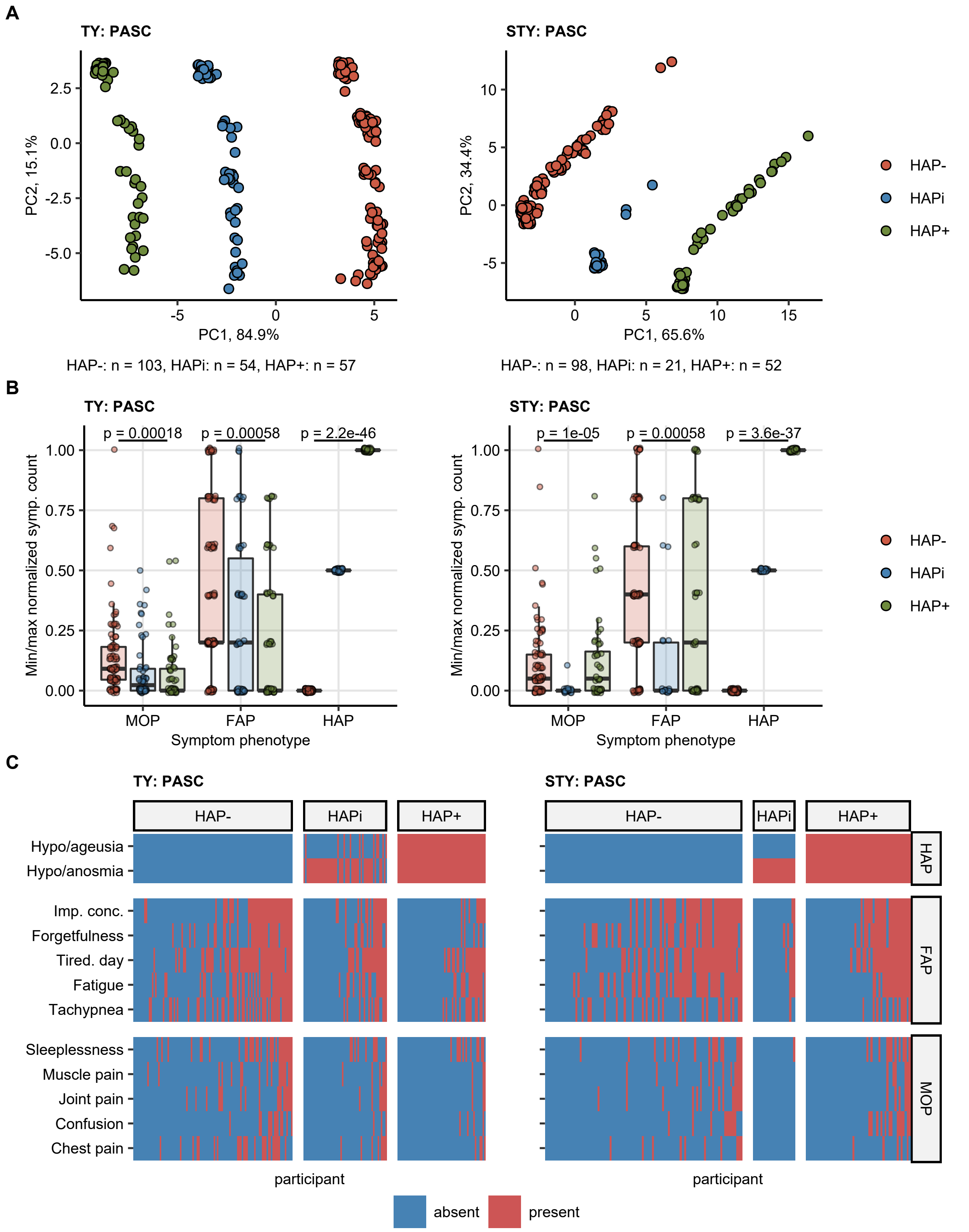


Figure S10: Subsets of PASC individuals defined by HAP, FAP and MOP phenotype symptoms.

Supplementary Figure S10. Subsets of PASC individuals defined by HAP, FAP and MOP

phenotype symptoms.

Hypo/anosmia-negative (HAP-), intermediate (HAPi) and high (HAP+) subsets of PASC individuals were defined in the training Tyrol (TY) cohort with Manhattan distance and DBSCAN clustering according to the counts of hypo/anosmia (HAP), fatigue (FAP) and multi-organ phenotype (MOP) symptoms. The subset assignment in the test South Tyrol (STY) cohort was done with k-nearest-neighbor label propagation algorithm. The analysis was conducted in the subsets of the study cohorts with the minimal observation time (SARS-CoV-2 test - survey) of 90 days.

(A) Two-dimensional principal component analysis (PCA) score plot with the PASC participant subset assignment. Percent variances associated with principal components (PC) are indicated in the plot axes. Numbers of subset individuals are indicated under the plots.

(B) Minimum/maximum-normalized counts of HAP, MOP and FAP symptoms in the PASC participant subsets. Differences between the participant subsets were investigated by Kruskal-Wallis test.

(C) Occurrence of the 10 most frequent HAP, FAP and MOP PASC symptoms (**Supplementary Figure S3**) in the PASC participant subsets presented as a heat map.

imp. conc.: impaired concentration, tired. day: tiredness at day.

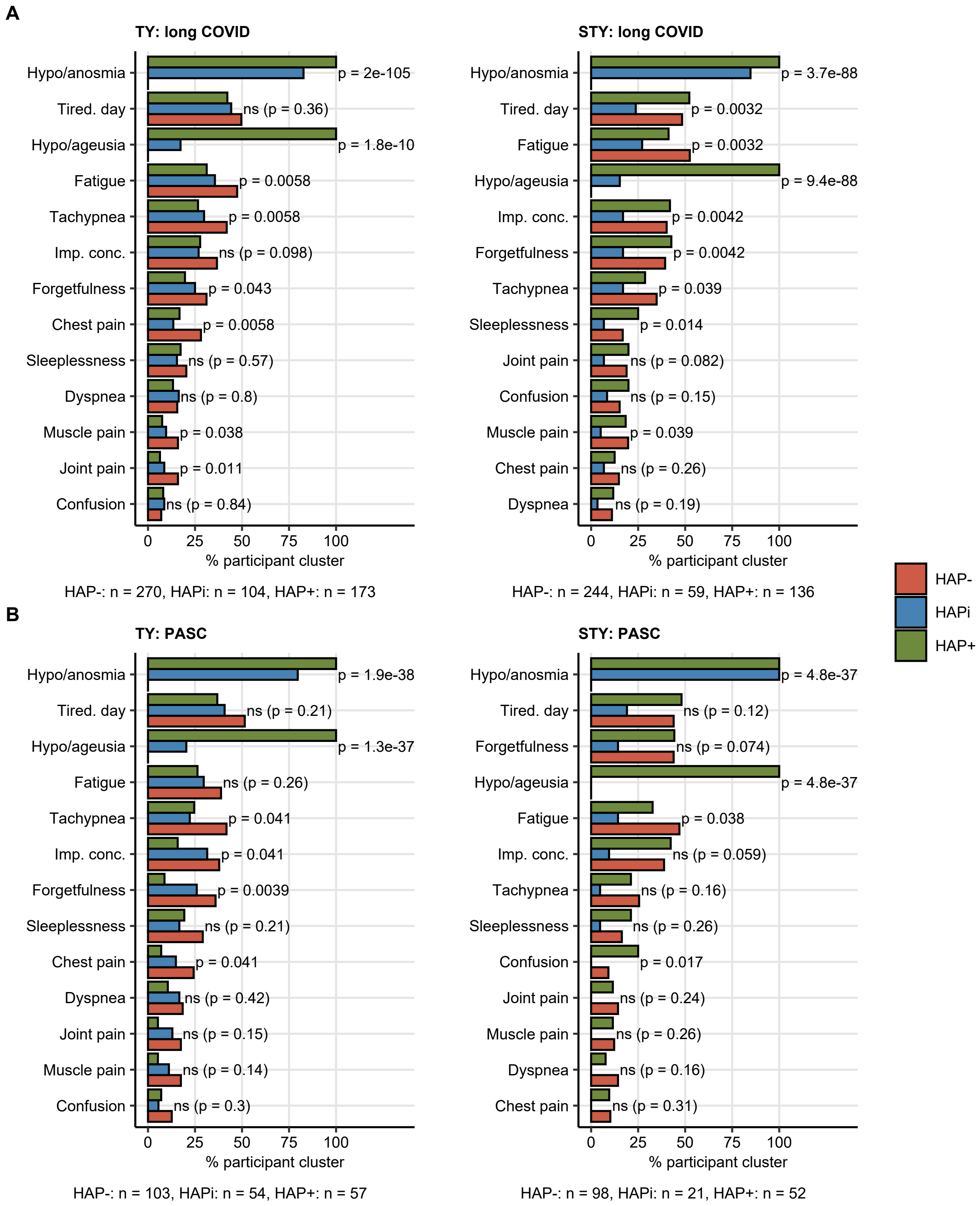


Figure S11: Subsets of PASC individuals defined by HAP, FAP and MOP phenotype symptoms.

Supplementary Figure S11. Frequency of the most frequent symptoms in the long COVID and PASC participant subsets.

Differences in frequency of the most frequent long COVID and PASC symptoms (**Supplementary Figure S3**) between the hypo/anosmia-negative (HAP-), intermediate (HAPi) and high (HAP+) subsets of long COVID (**A**) and PASC (**B**) individuals were investigated by χ^2 test and corrected for multiple comparisons with Benjamini-Hochberg method. Numbers of subset individuals are indicated under the plots.

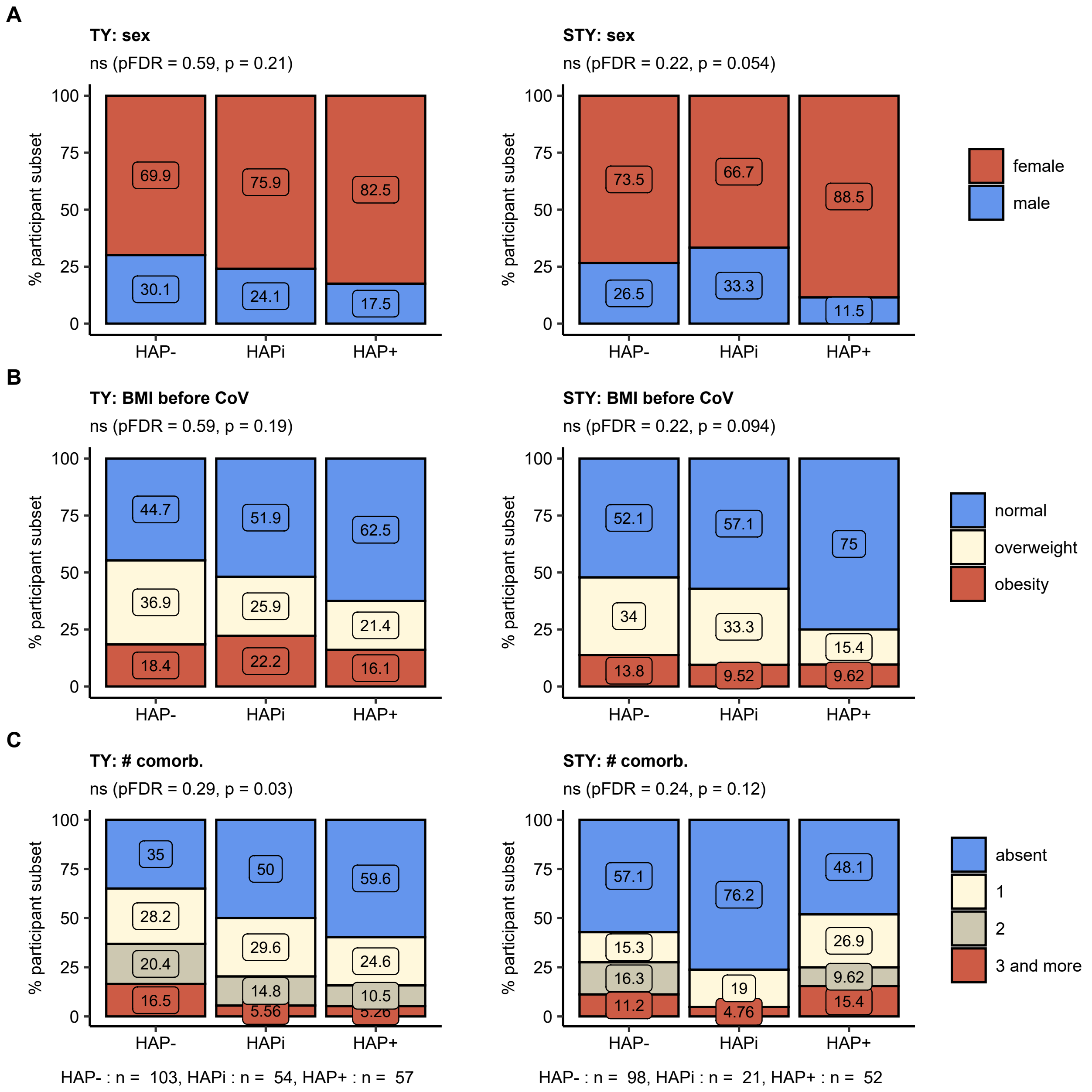


Figure S12: The most relevant demographic and clinical features of the PASC participant subsets.

Supplementary Figure S12. The most relevant demographic and clinical features of the PASC participant subsets.

Differences in demographic and clinical features (**Supplementary Table S5**) between the hypo/anosmia-negative (HAP-), intermediate (HAPi) and high (HAP+) subsets of long COVID individuals were investigated by χ^2 test. Comparison results for the most differentiating features: sex (**A**), body mass index class (**B**), number of comorbidities (**C**) and antibiotic therapy during acute COVID-19 (**D**) are presented. Raw and multiple testing-adjusted significance (pFDR) p values are presented in the plot captions. Numbers of subset individuals are indicated under the plots. TY: Tyrol, STY: South Tyrol.

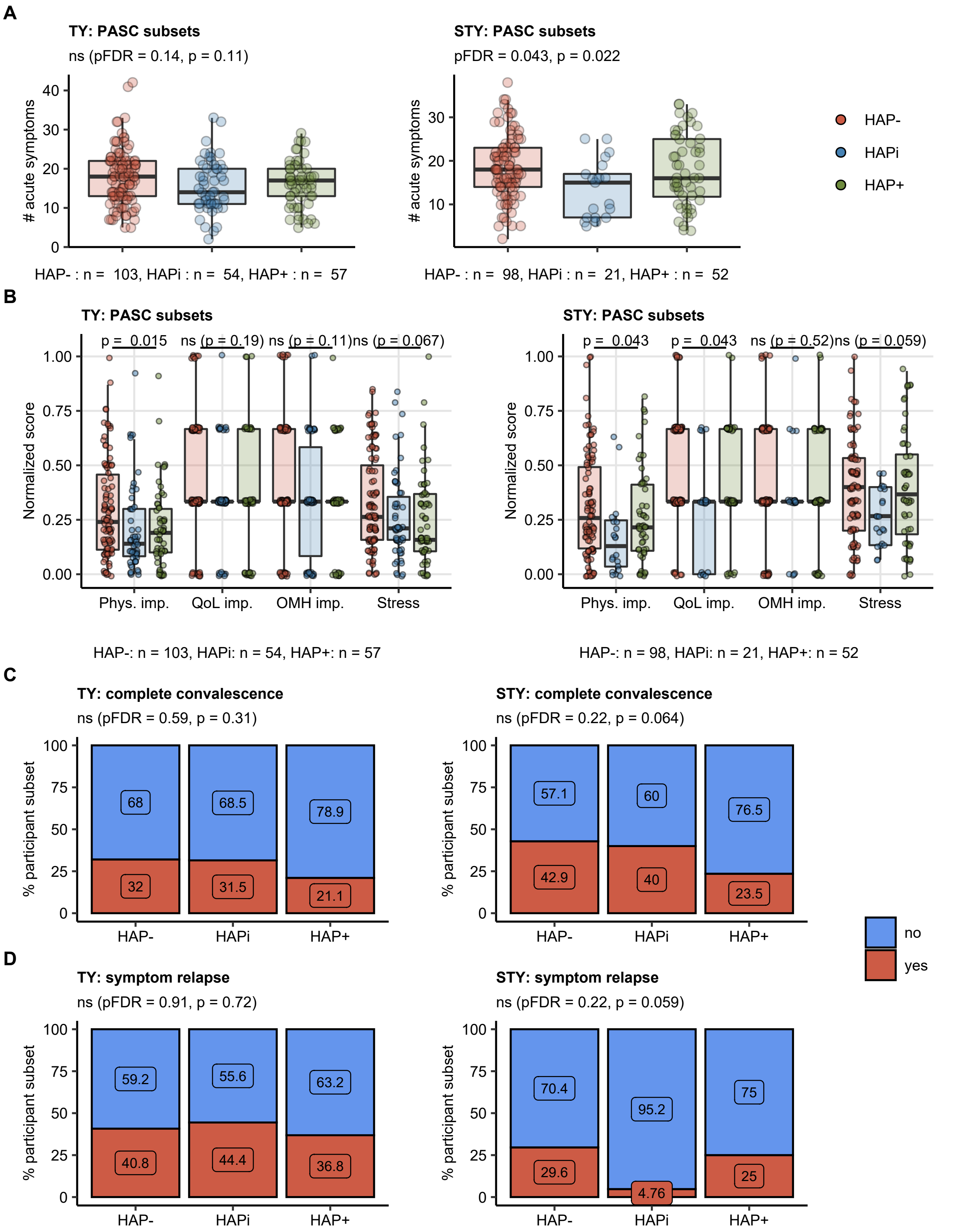


Figure S13: Acute symptom count, rating of physical, quality of life and mental impairment in the PASC participant subsets.

Supplementary Figure S13. Acute symptom count, rating of physical, quality of life and mental impairment in the PASC participant subsets.

(A) Numbers (#) of acute COVID-19 symptoms in the hypo/anosmia-negative (HAP-), intermediate (HAPi) and high (HAP+) subsets of PASC individuals. Statistical significance was assessed with Kruskal-Wallis test. Raw and multiple testing-adjusted significance (pFDR) p values are presented in the plot captions. Numbers of subset individuals are indicated under the plots.

(B) Minimum/maximum-normalized scores of physical performance (phys. imp), quality of life (QoL), overall mental health (OMH) impairment and stress in the subsets of PASC individuals. Statistical significance was assessed with Kruskal-Wallis test. Multiple testing-adjusted significance are presented in the plots.

(C - D) Frequencies of self-reported complete convalescence (**B**) and symptom relapse (**C**) in the PASC participant subsets. Statistical significance was assessed by χ^2 test. Raw and multiple testing-adjusted significance (pFDR) p values are presented in the plot captions.

TY: Tyrol, STY: South Tyrol.

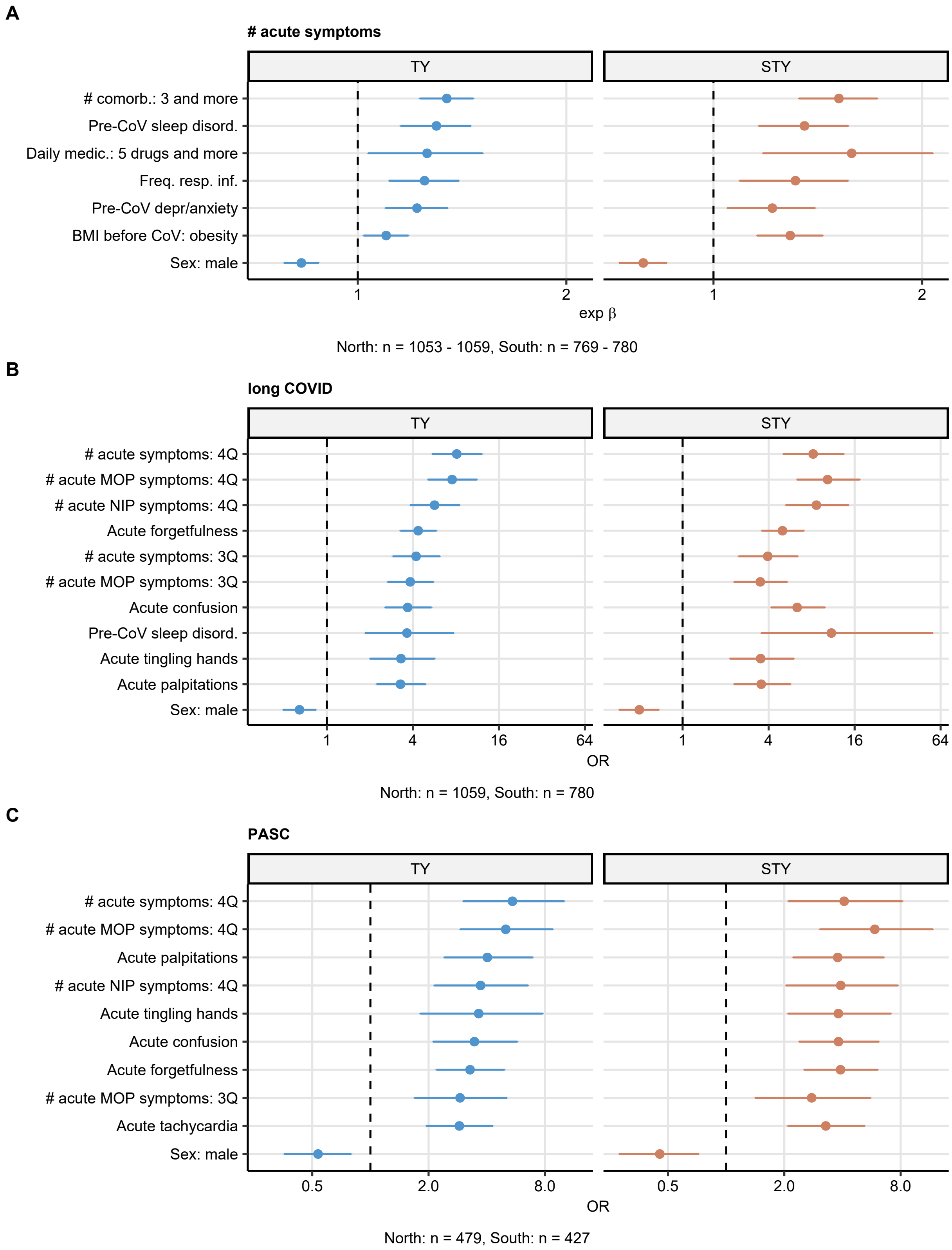


Figure S14: The major co-variates of acute COVID-19 symptom number, long COVID and PASC risk identified by univariable modeling.

Supplementary Figure S14. The major co-variates of acute COVID-19 symptom number, long COVID and PASC risk identified by univariable modeling.

Correlation of candidate factors (**Supplementary Table S7**) associated with the count of acute COVID-19 symptoms (**A**), risk of long COVID (**B**) and PASC (**C**) was investigated with a series of sex- and age-weighted ordinary Poisson (symptom number) or logistic (risk) models. Continuous observation time variable (SARS-CoV-2 test to completion interval) was included in the models as a confounder. Estimate significance was determined by Wald Z-test and corrected for multiple comparisons with Benjamini-Hochberg method. For the full list of significant factors, see: **Supplementary Table S9**. Estimate values (symptom counts: exponentiated β, risk: odds ratio/OR) with 95% CI for the ten strongest positive and negative co-variates are presented in Forest plots. Ranges of complete observations included in the models are shown under the plots.

TY: Tyrol, STY: South Tyrol, #: number, comorb.: comorbidities, medic.: medication, freq. resp. inf.: frequent respiratory infections ($>$ 2 per year), depr: depression, MOP: multi-organ phenotype, NIP: non-specific infection phenotype, disord.: disorder, 3Q, 4Q: 3rd and 4th symptom count quartile.

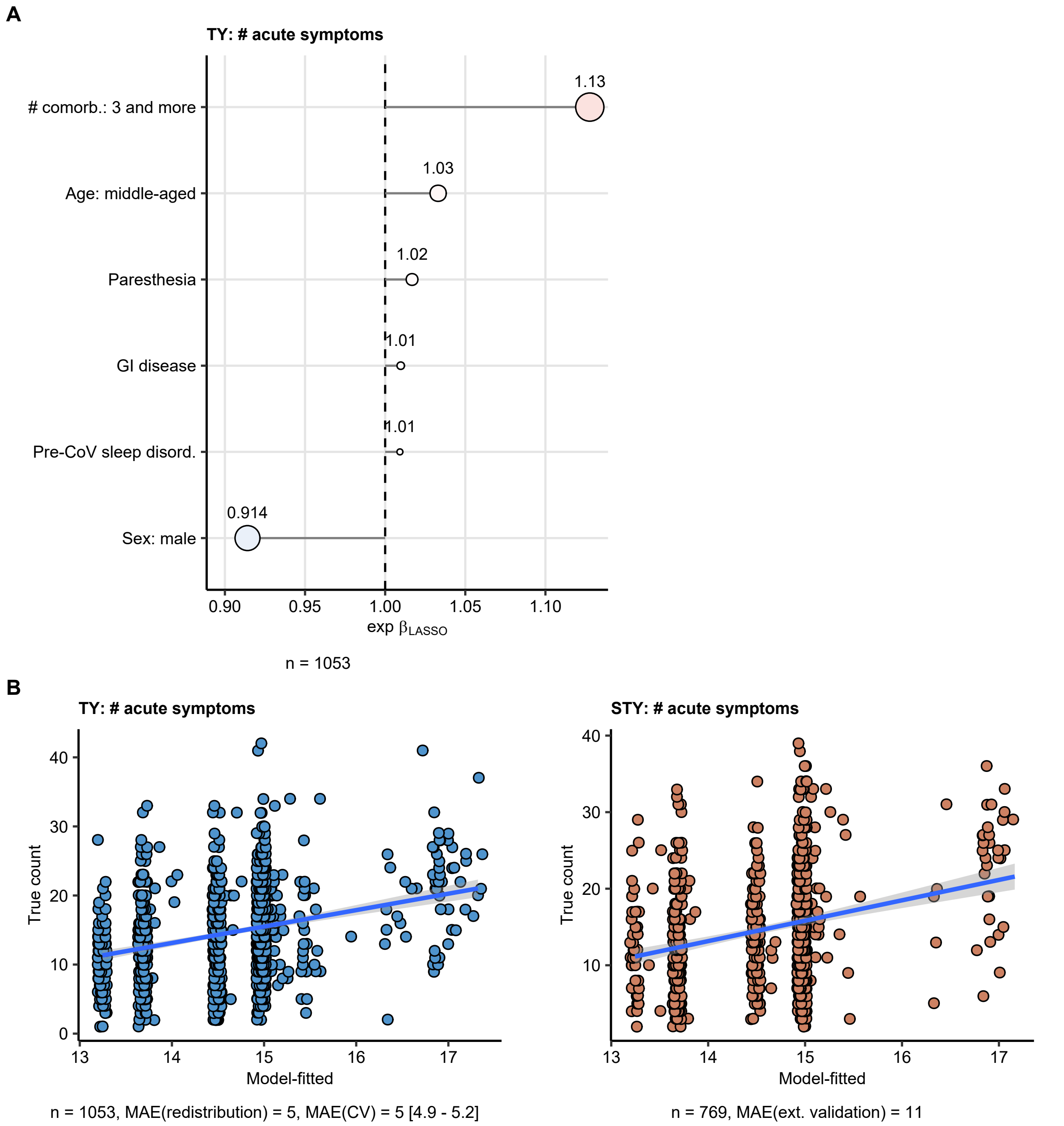


Figure S15: Identification of independent factors associated with acute COVID-19 symptom number by LASSO modeling.

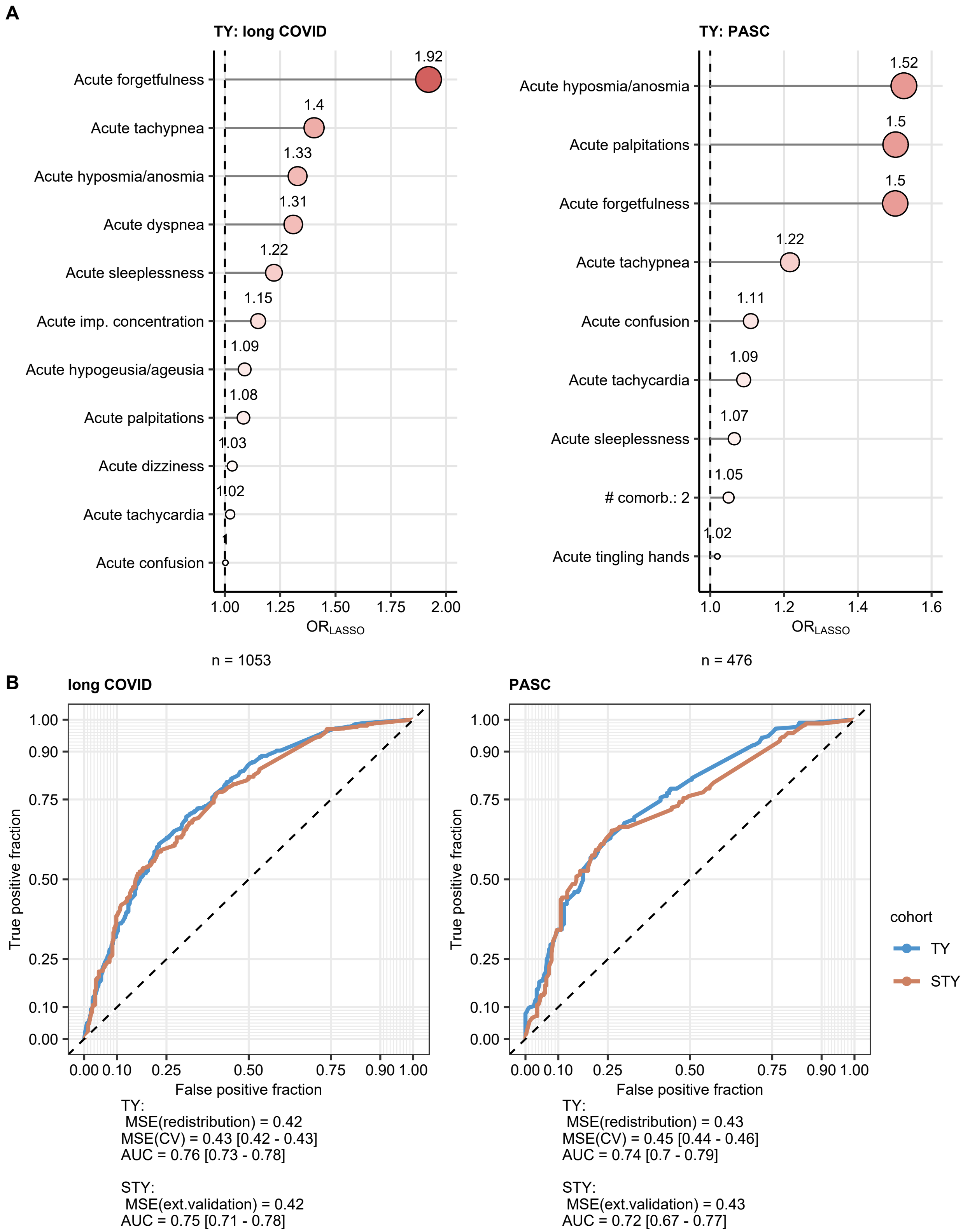
Supplementary Figure S15. Identification of independent factors associated with acute COVID-19 symptom number by LASSO modeling.

Correlation of candidate factors (**Supplementary Table S7**) with the count of acute COVID-19 symptoms was investigated by sex- and age-weighted LASSO (least absolute shrinkage and selection operator) Poisson regression in the training Tyrol (TY) cohort. Quality of model predictions was determined by assessment of redistribution error and 50-fold cross-validation (CV) in the Tyrol cohort and external validation in the South Tyrol (STY) collective.

(A) Values of non-zero β model estimates. Point size, fill and line length correspond with the exponentiated estimate value. The number of complete observations is indicated under the plot.

(B) Prediction of the acute COVID-19 symptom counts in the training TY and the test STY cohort. Numbers of complete observations, values of redistribution, cross-validation and external validation mean absolute errors (MAE) are indicated under the plots.

comorb.: comorbidities, #: number, disord.: disorder.



TY:
MSE(redistribution) = 0.42
MSE(CV) = 0.43 [0.42 - 0.43]
AUC = 0.76 [0.73 - 0.78]

STY:
MSE(ext.validation) = 0.42
AUC = 0.75 [0.71 - 0.78]

TY:
MSE(redistribution) = 0.43
MSE(CV) = 0.45 [0.44 - 0.46]
AUC = 0.74 [0.7 - 0.79]

STY:
MSE(ext.validation) = 0.43
AUC = 0.72 [0.67 - 0.77]

Figure S16: Identification of independent factors associated with long COVID and PASC risk by LASSO modeling.

Supplementary Figure S16. Identification of independent factors associated with long COVID and PASC risk by LASSO modeling.

Correlation of candidate factors (**Supplementary Table S7**) with the risk of long COVID and PASC was investigated by sex- and age-weighted LASSO (least absolute shrinkage and selection operator) logistic regression in the training Tyrol (TY) cohort. Quality of model predictions was determined by assessment of redistribution error and 50-fold cross-validation (CV) in the Tyrol cohort and external validation in the South Tyrol (STY) collective.

(A) Values of non-zero β model odds ratio (OR). Point size, fill and line length correspond with the OR value. Numbers of complete observations are indicated under the plot.

(B) Quality of prediction of long COVID and PASC risk in the training TY and the test STY cohort assessed by receiver-operator characteristic (ROC). AUC: area under the ROC curve.

comorb.: comorbidities, #: number, imp.: impaired.

References

1. Holzner B, Giesinger JM, Pinggera J, et al. The Computer-based Health Evaluation Software (CHES): A software for electronic patient-reported outcome monitoring. BMC Medical Informatics and Decision Making **2012**; 12. Available at: https://pubmed.ncbi.nlm.nih.gov/23140270/.

2. Sudre CH, Murray B, Varsavsky T, et al. Attributes and predictors of long COVID. Nature Medicine **2021**; 27. Available at: https://pubmed.ncbi.nlm.nih.gov/33692530/.

3. NICE. Overview | COVID-19 rapid guideline: managing the long-term effects of COVID-19 | Guidance | NICE.

4. Gräfe K, Zipfel S, Herzog W, Löwe B. Screening psychischer störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-D)". Ergebnisse der Deutschen validierungsstudie. Diagnostica **2004**; 50:171–181. Available at: https://econtent.hogrefe.com/doi/abs/10.1026/0012-1924.50.4.171.

5. Hüfner K, Tymoszuk P, Ausserhofer D, et al. Who is at risk of poor mental health following COVID-19 outpatient management? medRxiv **2021**;:2021.09.22.21263949. Available at: https://doi.org/10.1101/2021.09.22.21263949.

6. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. Journal of Open Source Software **2019**; 4:1686.

7. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 1st ed. New York: Springer-Verlag, 2016. Available at: https://ggplot2.tidyverse.org.

8. Wilke CO. Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. 1st ed. Sebastopol: O'Reilly Media, 2019: 389.

9. Sachs MC. Plotroc: A tool for plotting ROC curves. Journal of Statistical Software **2017**; 79:1–19. Available at: https://www.jstatsoft.org/index.php/jss/article/view/v079c02/v79c02.pdf%20https://www.jstatsoft.org/index.php/jss/article/view/v079c02.

10. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) **1995**; 57:289–300.

11. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. Journal of Statistical Software **2015**; 67:1–48. Available at: http://arxiv.org/abs/1406.5823.

12. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software **2017**; 82:1–26.

13. Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation. In: Society for industrial and applied mathematics - 8th siam international conference on data mining 2008, proceedings in applied mathematics 130. 2008: 243–254. Available at: https://experts.umn.edu/en/publications/similarity-measures-for-categorical-data-a-comparative-evaluation.

14. Schubert E, Rousseeuw PJ. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer, 2019: 171–187. Available at: https://link.springer.com/chapter/10.1007/978-3-030-32047-8_16.

15. Croux C, Filzmoser P, Oliveira MR. Algorithms for Projection-Pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems **2007**; 87:218–225.

16. Todorov V, Filzmoser P. Comparing classical and robust sparse PCA. In: Advances in intelligent systems and computing. Springer Verlag, 2013: 283–291. Available at: https://link.springer.com/chapter/10.1007/978-3-642-33042-1_31.

17. Drost H-G. Philentropy: Information Theory and Distance Quantification with R. Journal of Open Source Software **2018**; 3:765. Available at: https://doi.org/10.21105/joss.00765.

18. Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd international conference on knowledge discovery and data mining. 1996: 226–231. Available at: www.aaai.org.

19. Hahsler M, Piekenbrock M, Doran D. Dbscan: Fast density-based clustering with R. Journal of Statistical Software **2019**; 91:1–30. Available at: https://www.jstatsoft.org/index.php/jss/article/view/v091i01.

20. Belyadi H, Haghighat A, Nguyen H, Guerin AJ. IOP Conference Series: Earth and Environmental Science Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra Related content EPS conference comes to London-EPS rewards quasiparticle research-EP. IOP Conf Ser: Earth Environ Sci **2016**; 31.

21. Glennan T, Leckie C, Erfani SM. Improved Classification of Known and Unknown Network Traffic Flows Using Semi-supervised Machine Learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **2016**; 9723:493–501. Available at: https://link.springer.com/chapter/10.1007/978-3-319-40367-0_33.

22. Lelis L, Sander J. Semi-supervised density-based clustering. Proceedings - IEEE International Conference on Data Mining, ICDM **2009**;:842–847.

23. Leng M, Wang J, Cheng J, Zhou H, Chen X. Adaptive semi-supervised clustering algorithm with label propagation. Journal of Software Engineering **2014**; 8:14–22.

24. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) **1996**; 58:267–288. Available at: http://www.jstor.org/stable/2346178.

25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **2010**; 33:1–22. Available at: https://www.jstatsoft.org/index.php/jss/article/view/v033i01/v33i01.pdf%20https://www.jstatsoft.org/index.php/jss/article/view/v033i01.

26. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software **2014**; 61:1–36. Available at: https://www.jstatsoft.org/index.php/jss/article/view/v061i08.