UNIVERSITY OF GOTHENBURG

SCHOOL OF PUBLIC HEALTH AND COMMUNITY MEDICINE

Self-reported symptoms and their relation to COVID-19 infection and its severity

A Swedish pilot study

Anna Grimby-Ekman[1], Rode Grönkvist[1], Maria F. Gomez[2], Carole Sudre[4,3]

[1] School of Public Health and Community Medicine, Institute of Medicine, Gothenburg University, Sweden, [2]Department of Clinical Sciences, Lund University Diabetes Centre, Lund University, Sweden, [3] King's College London, London, UK; MRC Unit for Lifelong Health and Ageing, Department of Population Health Sciences and Centre for Medical Image Computing, Department of Computer Science

November 2023

School of Public Health and Community Medicine
Institute of Medicine
University of Gothenburg
Gothenburg, Sweden

Frontpage picture from Pixabay.

ISBN 978-91-527-2813-0

Content

Summary 4
Background 6
Materials and methods 8
Symptom period definition 8
Statistical analysis 10
Rasch analysis 10
Results 12
Rasch analysis on 5-day periods 14
Model 1 14
Model 2 15
Model 3 15
Model 4 17
Model 5 17
Model 6 18
Model 7 18
Unidimensionality 21
Differential item functioning 21
Sensitivity analysis on the subset of data 23
Validation of models 24
Model 3 24
Model 7 28
Discussion 33
References 36

Summary

On March 11th, 2020, the World Health Organization officially recognized COVID-19 as a global pandemic. Health systems worldwide were overwhelmed, in part due to the variety of new variants that began to emerge, presenting diverse symptom profiles and levels of infectiousness.

This study aimed to develop a predictive scale based on Rasch analysis, using symptoms reported by individuals with positive PCR tests. Additionally, the project sought to evaluate the scale's effectiveness in screening for positive PCR tests among those tested and in predicting hospitalization among those who tested positive.

Data were obtained from the COVID Symptom Study smartphone application in the United Kingdom, United States of America, and Sweden, with a focus on Swedish data from the COVID Symptom Study Sweden. Early symptoms, within the first 5 days, were of particular interest for early prediction of COVID-19 infection severity.

A Rasch analysis was used to investigate whether the symptoms could form a measurement scale related to COVID-19 infection. This could indicate the importance of the symptoms in regard to severity of the infection and regarding predictability. A low location of the symptoms on the the scale indicates that they are common and could indicate low severity. A high scale location would then indicate more rare symptoms that could be connected to more severe infection. Logistic regression was used to predict positive PCR tests among individuals who underwent PCR testing, as well as hospitalization among those with positive PCR tests.
The scale's fit to the Rasch model was moderate, its predictive ability for hospitalization among individuals with positive PCR tests was acceptable, as indicated by an area under the curve (AUC) of 0.7 (Mandrekar, 2010), but maybe not clinically useful (Fan et al., 2006).
However, the representation of COVID-19 cases in the Swedish data during the first half of 2020 was limited to more severe cases, primarily reflecting individuals with severe symptoms.

In conclusion, symptom clustering holds promise in understanding patterns in COVID-19 symptoms and could serve as a valuable screening tool for identifying severe cases. Further research, particularly focusing on predictive models and comparative analyses, is necessary to fully understand these symptom patterns and their practical applications. Our findings indicate that predicting COVID-19 severity is feasible, making continued research in this field imperative.

Acknowledgement: Adlerbert Research Foundation and Head Office has financially supported the work. We extend our gratitude to The COVID Symptom Study Sweden (CSSS) group and the COVIDX consortium for generously sharing their project data.

The COVID Symptom Study Sweden received financial support from the Swedish Heart-Lung Foundation (20190470, 20140776), the Swedish Research Council (EXODIAB, 2009-1039; 2014-03529), the European Commission (ERC-2015-CoG - 681742 NASCENT), eSSENCE@LU 8:8 (eSSENCE - The e-Science Collaboration), the Swedish Foundation for Strategic Research (LUDC-IRC, 15-0067), the Crafoord Foundation (20211011), and EUGLOHRIA (101017752).

ZOE Limited provided in kind support for all aspects of building, running, and supporting the COVID Symptom Study app used for data collection and service to all app users worldwide. ZOE Limited developed the app for data collection as a not-for-profit endeavour. None of the funding entities had any role in study design, data analysis, data interpretation, or the writing of this report.

Corresponding author:
Professr in Health Science Statistics
Anna Grimby-Ekman
Anna.ekman@gu.se

Background

Coronavirus disease 2019 (COVID-19) was reported as an epidemic in Wuhan China on December 31st, 2019 and declared a global pandemic by the World Health Organization (WHO) on March 11, 2020 (World Health Organization, 2020). COVID-19 continued to spread globally at an alarming rate. The health systems and health services in most countries around the world were stretched beyond expectation in 2020 and throughout 2021. Globally, waves of infection and new variants have emerged with varying symptoms, severity, and degree of infectiousness.

During the most intense period of the pandemic, there was a need to predict healthcare utilization, such as the potential number of care beds needed due to COVID-19 infections. Also, a need of effective diagnose or screening tools to identify positive COVID-19 cases and cases requiring hospitalization or ICU care. One approach was to focus on identifying COVID-19-related symptoms and their internal pattern in relation to infection and severity.

The COVID Symptom Study Sweden (CSSS) collected data on COVID-19 symptoms via a mobile app. The study collected data on self-reported symptoms and their relation to positive or negative COVID-19 PCR test, providing an opportunity to better understand the predictability of specific symptoms as indicators for possible COVID-19 infection and severity (Menni et al., 2020). Using data from the app, an unsupervised time series clustering, based on 14 symptoms, resulted in six symptom clusters (Sudre et al., 2021). The clusters were then used, together with demographic characteristics, to predict the need for respiratory support. The cluster analysis, together with additional analysis of the relation between different symptoms and their relationship with disease severity, holds significant promise in unraveling coexistence patterns of COVID-19 symptoms. However, additional research using predictive models and comparative analyses is necessary to fully understand these patterns between symptom and their practical applications. An alternative approach to cluster analysis is to use a Rasch analysis (Boone, 2016; Pallant & Tennant, 2007) to investigate whether the symptoms can form a measurement scale for the presence

and severity of COVID-19 infection. The Rasch analysis can also indicate the predictive ability of the symptoms on the scale.

If the data on symptoms, or a subset of the symptoms, fits into the Rasch model, it provides a scale with measurement properties (the distance between two scores is the same, regardless of where on the scale they lie). If the scale measures the presence or severity of COVID-19, it needs to be validated. Symptoms on the low part of the scale are common and have a high probability, while symptoms on the higher part of the scale are less common, with a "probability hierarchy" between the symptoms. If this pattern is not present in the symptoms, they will not fit into a Rasch model.

The primary objective of this study was to investigate whether symptoms reported by individuals with positive PCR tests could be utilized to develop a scale based on Rasch analysis. Additionally, we aimed to evaluate the screening capabilities of this scale for identifying positive PCR tests among individuals who underwent PCR testing, as well as predicting hospitalization among those with positive PCR tests.

Materials and methods

For this study, data were retrieved from the COVID Symptom Study (Kennedy et al., 2022). This is a population-based study with daily reports of symptoms from users of the COVID Symptom Study smartphone application, including data from the United Kingdom, the United States of America, and Sweden. Only data from users in Sweden were included in this study. The start of the study period was two weeks after the launch date in each country, to account for potentially unstable data close to launch. (Launch date in Sweden: 29 April 2020). Using data from the CSSS group was approved within the already existing ethical approval for CSSS (DNR 2020-01803 and supplements 2020-04006, 2020-04145, 2020-04451, 2020-07080, 2021-02316).

The overall inclusion criterion was that participants had at least one PCR test result (positive or negative COVID-19 infection) during the study period (N=11671 individuals). Of these individuals, 3 with a positive COVID-19 test were excluded due to marking other gender than man or woman; 3 individuals among those with a negative COVID-19 test were excluded (2 due to marking other gender than woman or man, and 1 due to missing value on gender). Since we investigated whether the scale properties were independent of sex or gender, sample sizes needed to be large enough in each sex group.

Symptom period definition

We investigated 19 symptoms, 14 of which were previously identified as important to predict COVID-19 infection (Sudre et al., 2021) and 4 additional symptoms (blisters on feet, red welts on face or lips, dizzy or lightheaded, and nausea). Symptom periods were defined as beginning on the first day with at least one symptom and ending on the day of one of the following: hospitalization or return from hospitalization, recovery, ending of acute symptom period (4-week duration of symptoms) or censure (the last day of reporting, when more than 5 consecutive days were missing). If five days or fewer were missing, a linear interpolation was used. Also, a symptom period was ongoing as long as that the symptoms was reported at

least three times over four days or more between the time their symptoms start and ends.

Only the first symptom period connected to a positive or negative COVID test (a test performed no earlier than seven days before and no later than seven days after the first day of the symptom period is included) was included in the study sample (N=11665; Figure 1).

For each symptom period, the symptoms during the first 5 days ('5-day period') were of special interest. This period indicated early symptoms, with the potential for early prediction of COVID-19 infection or severity of of COVID-19 infection. Hence, we hade two different registrations of symptoms. Those present in the total symptom period and those present in the first 5-day period.

The study sample then included 11665 individuals (Figure 1). We also analyzed a subset of data excluding those with lung disease, BMI over 30 and smokers. In this sub-set, 1719 individuals with positive COVID test were included (Figure 1).

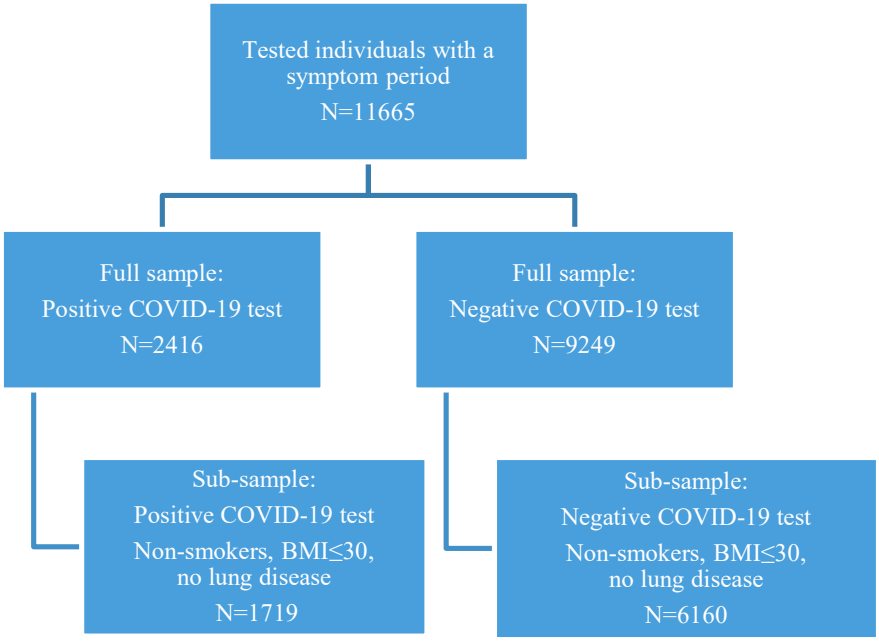


Figure 1. Study sample divided by those with positive or negative COVID-19 test result and the sub-samples, which excluded smokers, those with BMI≥30, and those with lung disease.

Statistical analysis

A Rasch analysis (Pallant and Tennant 2007, Boone 2016) was used to see if all symptoms or a subset of symptoms could represent a latent COVID-19 symptom score. The Rasch analysis was performed using the software RUMM2030 and the R-packages eRm (Mair & Hatzinger, 2007) and iarm (iarm: Item analysis in Rasch models; R package version 0.4.2). The Rasch analysis was performed on the data with symptoms during the 5-day periods, and only COVID-19-positive individuals were included.

To validate the symptom score as a predictor for being COVID-19 positive the severity of the COVID-19 infection (level of care needed), a logistic regression analysis was used, with the symptom score as a predictor. To predict positive COVID-19 cases, the total sample of those with either a negative or positive test result for COVID-19 infection were used. To predict the severity of COVID-19 infection, only COVID-19-positive individuals were included.

The prediction models were validated with receiver operating characteristic (ROC) analysis, presenting graphical display of the ROC and the precision-recall curves, and the area under the curve (AUC).

Rasch analysis

The Rasch model assumes an underlying latent variable that is not measurable. Using several measurable or assessable items a score is constructed, that is related to the latent variable. In this study, each symptom represented one of these items in the model. The latent variable is here assumed to measure COVID-19 presence or COVID-19 severity, whereby more symptoms indicate COVID-19 infection and severity. If the data fits the Rasch model, the Rasch scores on these items have fundamental measurement properties (the distance between two scores is the same regardless of where on the scale they lie). In this case, the latent variable is the severity of illness, with the idea that someone who is more severely ill with COVID-19 will present more symptoms.

Model fit, item fit, unidimensionality and differential item functioning (DIF) were evaluated. The reliability of the scale was evaluated with person separation reliability

(PSR), computed using the eRm package. A recommended threshold is PSR ≥0.7. Model fit was evaluated using the Andersen conditional likelihood ratio (CLR) test (Christensen, Kreiner, Mesbah 2012) as implemented in the iarm package (Mueller, 2022). Item fit was evaluated using conditional infit and outfit tests as well as bootstrapped outfit and infit statistics (Mueller, 2020), as implemented in the iarm package (Mueller, 2022). Item characteristic curves (ICC) were also created to visually evaluate item fit.

Item misfit was addressed by sequentially removing items that produced clear misfit. This process created a sequence of models for which model and item fits could be compared.

Uni-dimensionality was evaluated using the Martin-Löf test as implemented in the eRm package (Mair, P., & Hatzinger, R. 2007), in which items were divided into two hypothetical dimensions by performing principal component analysis (PCA) on item residuals in the model.

DIF was evaluated for the following personal factors: sex (male/female), lung disease status (yes/no), being a healthcare professional (no/yes, no direct contact with patients/direct contact with patients), smoker status (no/not currently/yes), age category (18-40/41-65/66+) and BMI category (<25/25-30/>30). DIF was evaluated in two steps: first on a model basis by performing the Andersen CLR test on the model, dividing it by each personal variable in turn, and secondly on an item basis by performing partial gamma tests for DIF on each combination of items and personal variables.

Results

The group with a positive COVID-19 test, compared to those with a negative test, included a slightly lower prevalence of women, health care professionals and individuals aged 18-40. Lung disease was slightly more common among those with a positive COVID-19 test.

Table 1. Background variables in the full dataset, stratified by postitive or negative COVID-19 test.

		Positive COVID-19 test N=2416		Negative COVID-19 test N=9249	
Variable		**Number**	**%**	**Number**	**%**
Gender	Female	1779	74	7277	79
	Male	637	26	1972	21
Lung disease	False	2167	90	7979	86
	True	249	10	1270	14
Healthcare professional	No	1671	69	6951	75
	Yes, interaction	578	24	1633	18
	Yes, no interaction	167	7	665	7
Smoker	Never	1583	66	5904	64
	Not currently	735	30	2841	31
	Yes	98	4	504	5
Age category	18-40	711	29	2932	32
	41-65	1559	65	5625	61
	Over 65	146	6	692	7
BMI category	Under 25	1116	46	4255	46
	25-30	854	35	3138	34
	Over 30	446	18	1856	20
Period end	Hospitalized	149	6	388	4
	Long Period	50	2	50	1
	Recovered	2112	87	8040	87
	Censored	105	4	771	8

The prevalence of symptoms in the study sample was computed both for the 5-day period and the total symptom period, stratified by positive or negative COVID-19 test (table 2). Several symptoms, most notably loss of smell, were much more common in the COVID-19 positive group. The symptom sore throat was more common in the COVID-19 negative group.

Table 2. Prevalence of symptoms over the 5-day periods and the total symptom periods. Study sample N=11665.

	COVID-19 positive N=2416				COVID-19 negative N=9249			
	5-day period		Total symptom period		5-day period		Total symptom period	
Symptom	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**
Fever	61	1475	64	1542	42	3901	44	4064
Loss of smell	57	1367	67	1630	15	1418	17	1528
Unusual muscle pains	33	788	35	846	19	1784	21	1914
Persistent cough	43	1027	47	1138	35	3262	38	3540
Fatigue	18	446	23	558	12	1141	14	1283
Shortness of breath	5	120	7	180	5	457	6	538
Diarrhea	24	591	30	736	20	1808	22	2018
Delirium	12	278	15	352	10	886	11	986
Skipped meals	32	782	37	895	22	2042	24	2194
Abdominal pain	19	464	23	551	17	1541	19	1716
Chest pain	26	623	30	730	20	1860	22	2020
Headache	76	1834	80	1925	70	6448	73	6733
Chills or shivers	30	728	32	780	26	2408	28	2571
Eye soreness	37	890	41	1000	35	3256	37	3452
Nausea	23	567	30	717	22	1993	24	2190
Dizzy or light-headed	41	991	48	1151	31	2835	33	3056
Red welts on face or lips	8	184	10	246	6	600	7	688
Blisters on feet	1	20	1	28	1	106	1	123
Sore throat	59	1415	61	1482	77	7117	78	7253

Rasch analysis on 5-day periods

A step-wise Rasch analysis was done on the 5-day periods, culminating in a final set of symptoms (items) included in the Rasch model. Each step in the procedure is presented below (summarized in Table 3), described as models 1-7.

Model 1

Model 1 included all symptoms in the dataset as items.

The Andersen CLR test was used to test the model fit of Model 1, p-value <0.0001, indicating the data did not fit the model (CLR 340.527, 18 degrees of freedom.). PSR was 0.648, which is a moderate level as >0.8 is desirable. Bootstrap infit/outfit tests indicated significant item misfit in both infit and outfit for the following items: loss of smell, unusual muscle pains, persistent cough, fatigue, delirium, abdominal pain, headache, nausea, dizzy or light-headed, and sore throat. ICC analysis showed that ‘loss of smell’ and ‘sore throat’ had much more severe misfit than other items. In both cases, the ICC indicated severe underdiscrimination.

PCA of residuals indicated potential deviation from unidimensionality, with one dimension consisting of the following items: fever, unusual muscle pains, persistent cough, headache, chills or shivers, and sore throat; and the other dimension consisting of remaining items. A Martin-Löf test of these dimensions showed strong significance, with a p-value of 8.26e-27.

As a result of this analysis, 'loss of smell’ and ‘sore throat’ were removed from the model.

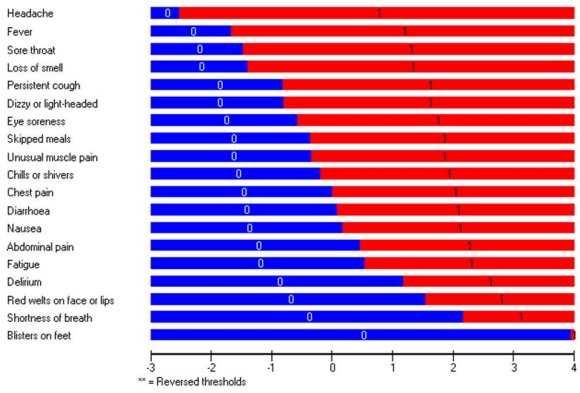


Figure 2. Full data set with symptoms over the 5-day period, showing no reversed thresholds in Model 1.

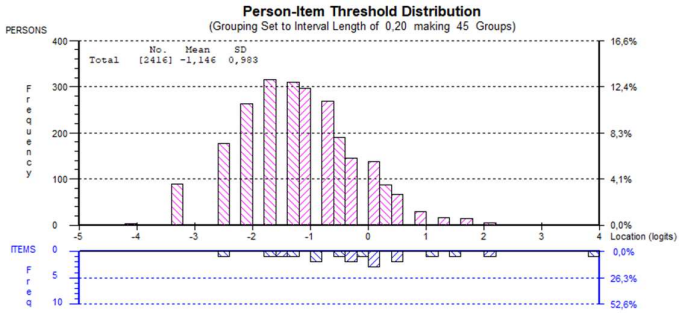


Figure 3. Full data set with symptoms over the 5-day period, showing item and person locations in Model 1.

Model 2

Model 2 included all symptoms except 'loss of smell' and 'sore throat'.

The Andersen CLR test was used to test the model fit of Model 2, $p<0.0001$, indicating the data did not fit the model (CLR 130.854, 16 d.f.), but showing greatly improved fit compared to model 1. PSR was 0.625. Bootstrap infit/outfit tests indicated significant item misfit in both infit and outfit for the following items: unusual muscle pains, persistent cough, diarrhoea, fatigue, delirium, nausea, dizzy or light-headed, and red welts on face or lips. ICC analysis showed that 'persistent cough' had much more severe misfit than other items.

As a result of this analysis, 'persistent cough' was removed from the model.

Model 3

Model 3 included all symptoms except 'loss of smell', 'sore throat', and 'persistent cough'.

The Andersen CLR test was used to test the model fit of Model 3,p-value=2.31e-11, indicating the data did not fit the model (CLR 82.616, 15 d.f.), but showing improved fit compared to Model 2.

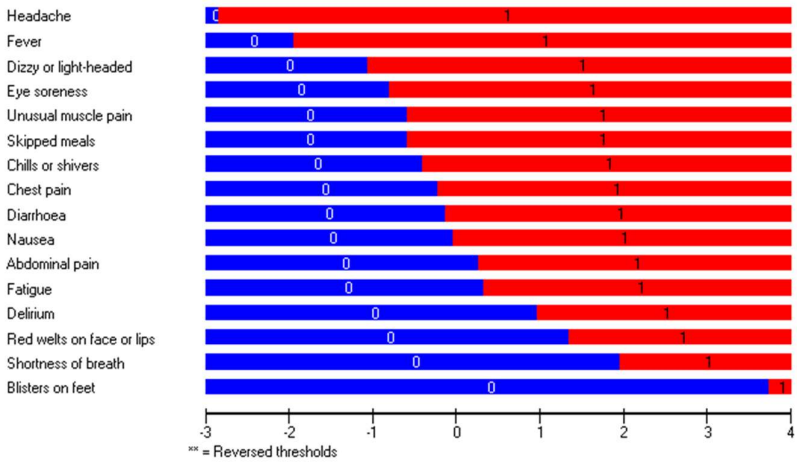


Figure 4. Full dataset with symptoms in the 5-day period, showing no reversed thresholds in Model 3.

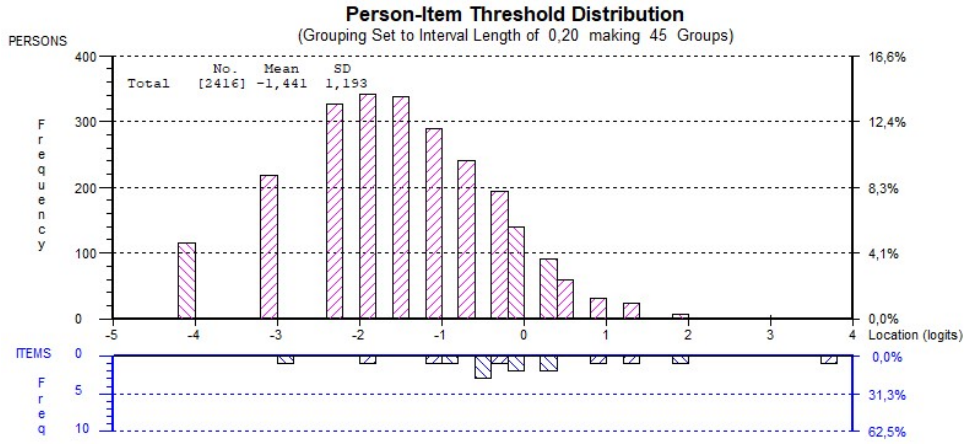


Figure 5. Full data set with symptoms in the 5-day period, showing item and person locations in Model 3.

PSR was 0.610. Bootstrap infit/outfit tests indicated significant item misfit in both infit and outfit for the the following items: fever, unusual muscle pains, diarrhoea, fatigue, delirium, nausea and chest pain. ICC analysis indicated similar misfit patterns for ‘fatigue’ and ‘diarrhoea’, and bootstrap infit/outfit tests showed these two symptoms to be most strongly misfit, with ‘diarrhoea’ having the strongest misfit. As a result of this analysis, ‘diarrhoea’ was removed from the model.

Model 4

Model 4 included all symptoms except fo the following: loss of smell, sore throat, persistent cough, and diarrhoea.

The Andersen CLR test was used to test the model fit of Model 4, p-value=3.20e-8, , indicating the data did not fit the model (CLR 63.196, 14 d.f.), but showing improved fit compared to Model 3. PSR was 0.601. Bootstrap infit/outfit tests indicated significant item misfit in both infit and outfit for the following items: fever, fatigue, chest pain, and red welts on face or lips. ICC analysis indicated similar misfit patterns for 'fatigue' as for 'diarrhoea' in the previous model, and bootstrap infit/outfit tests showed that 'fatigue' was strongly misfit.

As a result of this analysis, 'fatigue' was removed from the model.

Model 5

Model 5 included all symptoms except for the following: loss of smell, sore throat, persistent cough, diarrhoea, and fatigue.

The Andersen CLR test was used to test the model fit of Model 5, p-value=1.50e-5, indicating the data did not fit the model (CLR 45.869, 13 d.f.), but showing improved fit compared to Model 4. PSR was 0.574. This drop in PSR compared to the previous model was notably greater than for earlier models. Bootstrap infit/outfit tests did not indicate significant item misfit in both infit and outfit for any items. However, the item 'unusual muscle pains' displayed significant infit misfit and the items 'abdominal pain', 'chest pain', 'nausea', and 'red welts on face or lips' displayed significant outfit misfit. ICC analysis indicated that the items with significant outfit misfit were particularly misfit close to the top of the ICC, whereas the item with significant infit misfit showed a small tendency towards over-discrimination.

As a result of this analysis, attempts were made to formulate new models by removing one of the items displaying misfit at a time. Of these models, the model with 'red welts on face or lips' removed displayed an improved CLR test statistic and PSR, whereas removing any other symptom resulted in a worse model than when the symptom was included. Thus, the 'red welts on face or lips' was removed.

Model 6

Model 6 included all symptoms except the following: loss of smell, sore throat, persistent cough, diarrhoea, fatigue, and red welts on face or lips.

The Andersen CLR test was used to test the model fit of Model 6, p-value=0.004, indicating data do not fit the model (CLR 28.86, 12 d.f.), but showing improved fit compared to Model 5. PSR was 0.575. This model improved in both PSR and CLR as compared to the previous model, whereas earlier models improved in one but worsened in the other. Bootstrap infit/outfit tests did not indicate significant item misfit in either infit or outfit for any items. However, the items ‘unusual muscle pains’ and ‘eye soreness’ displayed significant infit misfit, and the item ‘chest pain’ displayed significant outfit misfit. ICC analysis gave no clear reason to remove any one item in comparison to any other.

As a result of this analysis, attempts were made to formulate new models by removing one of the items displaying misfit at a time. Of these models, the model with ‘chest pain’ removed displayed an improved CLR test statistic but worsened PSR, whereas removing any other symptom resulted in a worse model both by CLR test statistic and PSR, compared to not removing the symptom. Thus, the item ‘chest pain’ was removed.

Model 7

Model 7 included all symptoms except the following: loss of smell, sore throat, persistent cough, diarrhoea, fatigue, red welts on face or lips, and chest pain.

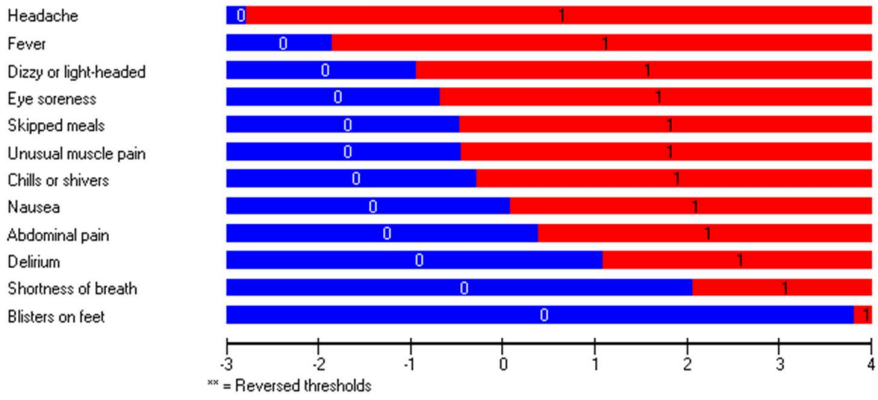


Figure 6. Showing no reversed thresholds in model 7. Full data set with symptoms in the 5-day period.

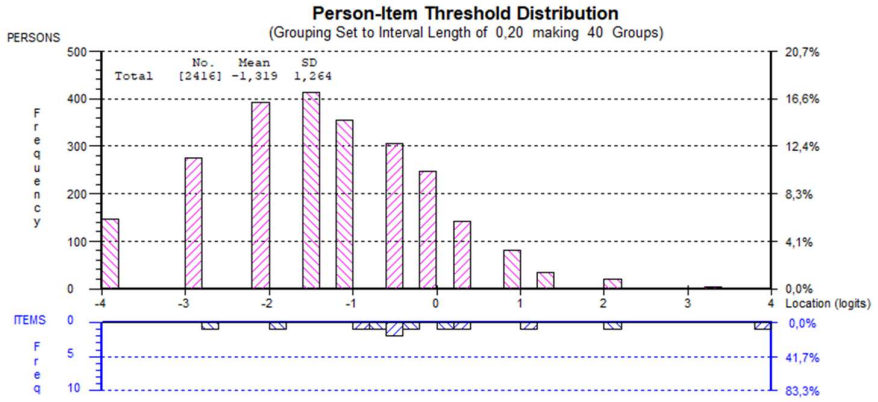


Figure 7. Showing item and person locations in model 7. Full data set with symptoms in the 5-day period.

The Andersen CLR test was used to test the model fit of Model 7, p-value=0.176, indicating the data fit the model (CLR 15.139, 11 d.f.). This made it the first model for which the data fit, as indicated by the Andersen CLR test. PSR was 0.556. Bootstrap infit/outfit tests did not indicate significant item misfit in either infit or outfit for any items.

Table 3. Summary of the step-wise procedure for Rasch analysis of the 5-day periods, full dataset

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Model fit: Andersen CLR test	p<0.0001	p<0.0001	p<0.0001	p<0.0001	p<0.0001	p=0.004	p=0.176
Pearson Separation Reliability (PSR)	0.648	0.625	0.610	0.601	0.574	0.575	0.556
Bootstrap overall item infit/outfit tests	*Misfit in infit and outfit:* Loss of smell Unusual muscle pains Persistent cough Fatigue Delirium Abdominal pain Headache Nausea Dizzy or light-headed Sore throat	*Misfit in infit and outfit:* Unusual muscle pains Persistent cough Diarrhoea Fatigue Delirium Nausea Dizzy or light-headed Red welts on face or lips	*Misfit in infit and outfit:* Fever Unusual muscle pains Diarrhoea Fatigue Delirium Nausea Chest pain	*Misfit in infit and outfit:* Fever Fatigue Chest pain Red welts on face or lips	*Infit misfit:* Unusual muscle pains *Outfit misfit:* Abdominal pain Chest pain Nausea Red welts on face or lips	*Infit misfit:* Unusual muscle pains Eye soreness *Outfit misfit:* Chest pain	No test indicated significant item misfit in either infit or outfit for any items.
Item characteristic curves (ICC)	Loss of smell and Sore throat	Persistent cough	Fatigue and Diarrhoea	Fatigue			
Dropped items at end of analysis	Loss of smell Sore throat	Persistent cough	Diarrhoea	Fatigue	Red welts on face or lips	Chest pain	

Unidimensionality

Investigation of unidimensionality showed consistent evidence of bidimensionality. Martin-Löf tests were found to be strongly significant for all models. The dimensions found by investigating residual principal components were consistent across models. One dimension, Dimension 1, consisted of the following sypmtoms: fever, unusual muscle pains, persistent cough, headache, chills or shivers, and sore throat (or whatever subset of these symptoms were present in the model). The other dimension, Dimension 2, consisted of the remaining symptoms. A possible interpretation of these dimensions is that the first consisted of symptoms typically associated with influenza, whereas the second consisted of all other symptoms.

To investigate whether either of these dimensions could be used in a Rasch model, the items from both dimensions were modeled separately as Rasch models, after first removing the items ‘loss of smell’, 'sore throat’, and ‘persistent cough’. Hence, the symptoms included in Model 3 above here were split into the two dimensions. The data showed a much worse fit both for the model for Dimension 1 and for the model for Dimension 2, compared to Model 3. The influenza symptom model (Dimension 1) had a CLR test p-value of 0.013 (10.768, 3 d.f.) and a PSR of -0.218. The other symptom model (Dimension 2) had a CLR test p-value of 0 (CLR 104.687, 11 d.f.) and a PSR of 0.359. Thus, the Model 3 was preferred despite some evidence of multidimensionality in its items.

Differential item functioning

DIF was first evaluated for Model 1, using every item. Here, DIF was found for 'fever’, ‘loss of smell’, ‘chills or shivers’, and ‘nausea’. At this point, the DIF of these items was considered a smaller issue than the large misfit for ‘loss of smell’, ‘sore throat’, and ‘persistent cough’ items, and so these were removed first.

Next, Model 3 was examined for DIF. DIF was found for the following items: fever, shortness of breath, delirium, chills or shivers, and nausea. The same items displayed DIF in Model 7. Since none of these items had infit or outfit tests for misfit, it was considered that their DIF was not a major issue for the model, and these items were not adjusted for differential item functioning.

Table 4 presents all items with significant DIF in Model 7 according to a partial gamma test, as well as the estimated Goodman-Kruskal gamma of the item-DIF factor combination.

Table 4. Significant DIF in Model 7 for the outcome of symptoms in the first 5-day period.

Item	DIF factor	Gamma
Fever	Gender	0.35
	Healthcare professional	-0.18
	Age category	0.32
Shortness of breath	Lung disease	0.50
Delirium	Age category	-0.29
Chills or shivers	Gender	0.38
Nausea	Gender	-0.43
	Healthcare professional	0.21

Table 5. Full data set with symptoms over the 5-day period. The mean locations for background variables were estimated for the full dataset with symptoms in the 5-day period in model 7.

Personal variable	Variable categories	n	Mean of logit	Standard deviation of logit
Total	-	2416	-1.319	1.264
Gender	Female	1779	-1.280	1.27
	Male	637	-1.427	1.24
Lung disease	False	2167	-1.359	1.25
	True	249	-0.967	1.37
Healthcare professional	No	1671	-1.394	1.26
	Yes, interaction	578	-1.162	1.27
	Yes, no interaction	167	-1.107	1.25
Smoker	Never	1583	-1.407	1.25
	Not currently	735	-1.187	1.27
	Yes	98	-0.893	1.30
Age category	18-40	711	-1.276	1.30
	41-65	1559	-1.319	1.25
	Over 65	146	-1.533	1.23
BMI category	Under 25	1116	-1.460	1.26
	25-30	854	-1.247	1.27
	Over 30	446	-1.105	1.22
Period end	Censored	105	-1.521	1.40
	Hospitalized	149	-0.604	1.43
	Long period	50	-1.321	1.28
	Recovered	2112	-1.359	1.23

Sensitivity analysis on the subset of data

As a sensitivity analysis, a Rasch analysis was done on the subset of data that excluded smokers, those with BMI≥30, and those with lung disease.

Table 6. Prevalence of symptoms in the 5-day periods and in the total symptom periods. Sub-sample N = 7879.

	COVID-19 positive N = 1719				COVID-19 negative N = 6160			
	5-day period		Total symptom period		5-day period		Total symptom period	
Symptom	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**
Fever	59	1022	62	1063	40	2485	42	2578
Loss of smell	56	962	67	1152	15	904	16	972
Unusual muscle pains	31	529	33	566	17	1027	18	1105
Persistent cough	40	689	44	761	33	2023	36	2200
Fatigue	17	293	21	365	11	663	12	745
Shortness of breath	4	64	6	97	4	217	4	257
Diarrhea	23	388	28	473	17	1056	19	1178
Delirium	11	181	13	218	9	524	9	581
Skipped meals	30	519	34	593	20	1210	21	1302
Abdominal pain	18	301	21	354	16	973	17	1073
Chest pain	24	420	28	479	19	1195	21	1286
Headache	74	1278	78	1349	68	4200	71	4390
Chills or shivers	28	485	30	523	25	1526	26	1629
Eye soreness	35	599	39	671	33	2004	34	2122
Nausea	22	380	28	476	20	1247	22	1362
Dizzy or light-headed	39	665	45	770	29	1758	31	1901
Red welts on face or lips	8	133	10	173	6	344	6	399
Blisters on feet	1	14	1	18	1	60	1	69
Sore throat	59	1011	61	1056	78	4788	79	4874

Table 7. Full data set with symptoms in the 5-day period. The mean locations for background variables estimated for the subset, with symptoms in the 5-day period in model 7.

Personal variable	Variable categories	n	Mean of logit	Standard deviation of logit
Total	-	1719	-1.425	1.254
Gender	Female	1241	-1.384	1.26
	Male	478	-1.532	1.23
Lung disease	False	1719	-1.425	1.254
	True	0	-	-
Healthcare professional	No	1223	-1.482	1.26
	Yes, interaction	389	-1.271	1.24
	Yes, no interaction	107	-1.335	1.20
Smoker	Never	1219	-1.463	1.24
	Not currently	500	-1.332	1.28
	Yes	0	-	-
Age category	18-40	497	-1.391	1.28
	41-65	1114	-1.421	1.24
	Over 65	108	-1.632	1.29
BMI category	Under 25	989	-1.509	1.24
	25-30	730	-1.312	1.26
	Over 30	0	-	-
Period end	Censored	70	-1.585	1.37
	Hospitalized	94	-0.900	1.43
	Long period	35	-1.462	1.37
	Recovered	1520	-1.449	1.23

Validation of models

Two validations were performed on Model 3 and Model 7 for the 5-day periods: one on the prediction of hospitalization and one on the prediction of COVID-19 infection.

Model 3

Predicting hospitalization

To predict hospitalization, only the individuals with a positive PCR test were included.

The graphical description of the location values and the prediction probabilities from the prediction mode, Figure 8, illustrates that both logits and probabilities were overlapping in the two groups of hospitalization or not.

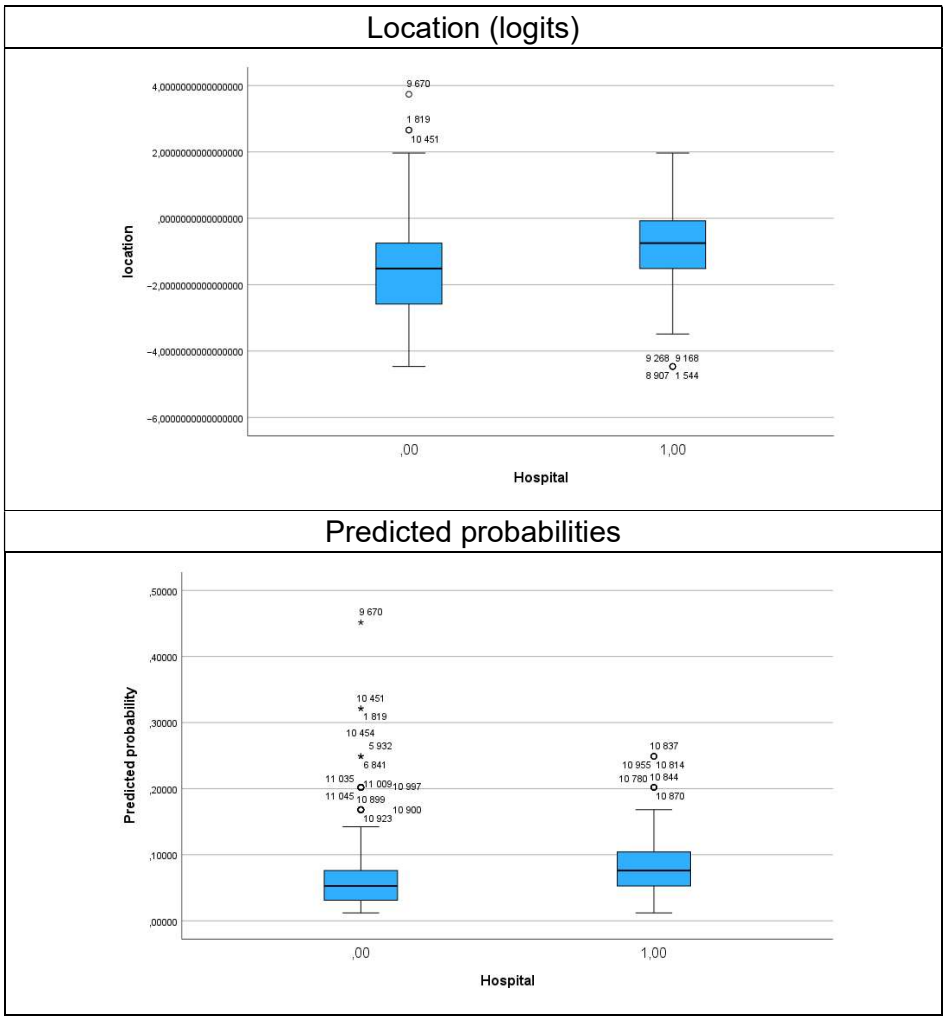


Figure 8. Model 3, box plots of logits (individual locations) from the Rasch analysis, and probability of being COVID-positive, estimated from the logistic regression.

In Figure 9, the ROC curve is presented, with an AUC of 0.67 (95% CI 0.623; 0.717). The precision-recall curve is also presented. Including sex and age in the logistic model only slightly improved the AUC.

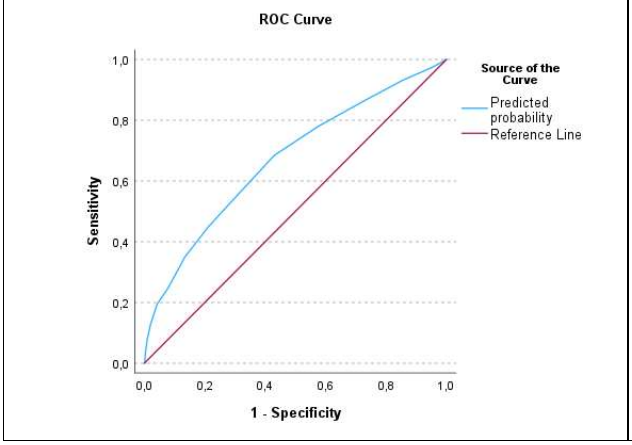
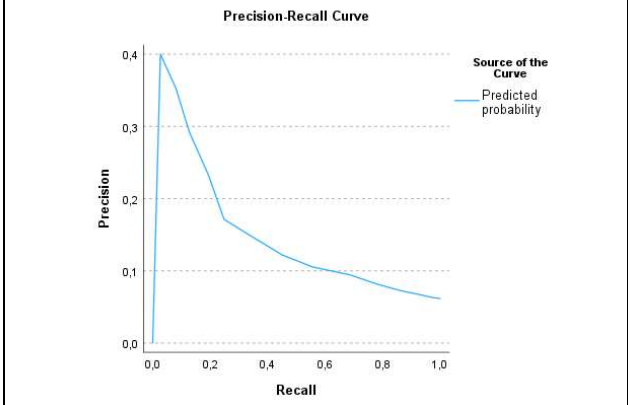
ROC curve	Precision-Recall Curve
	

Figure 9. Predicting hospitalization. Model 3, the ROC Curve and precision-recall curve (positive prediction value-sensitivity).

Predicting positive or negative PCR test result

To predict a positive PCR test result or not, the full dataset of individuals with either a positive or a negative PCR test was used.

A graphical description of the location values and the prediction probabilities from the prediction model is shown in Figure 10.

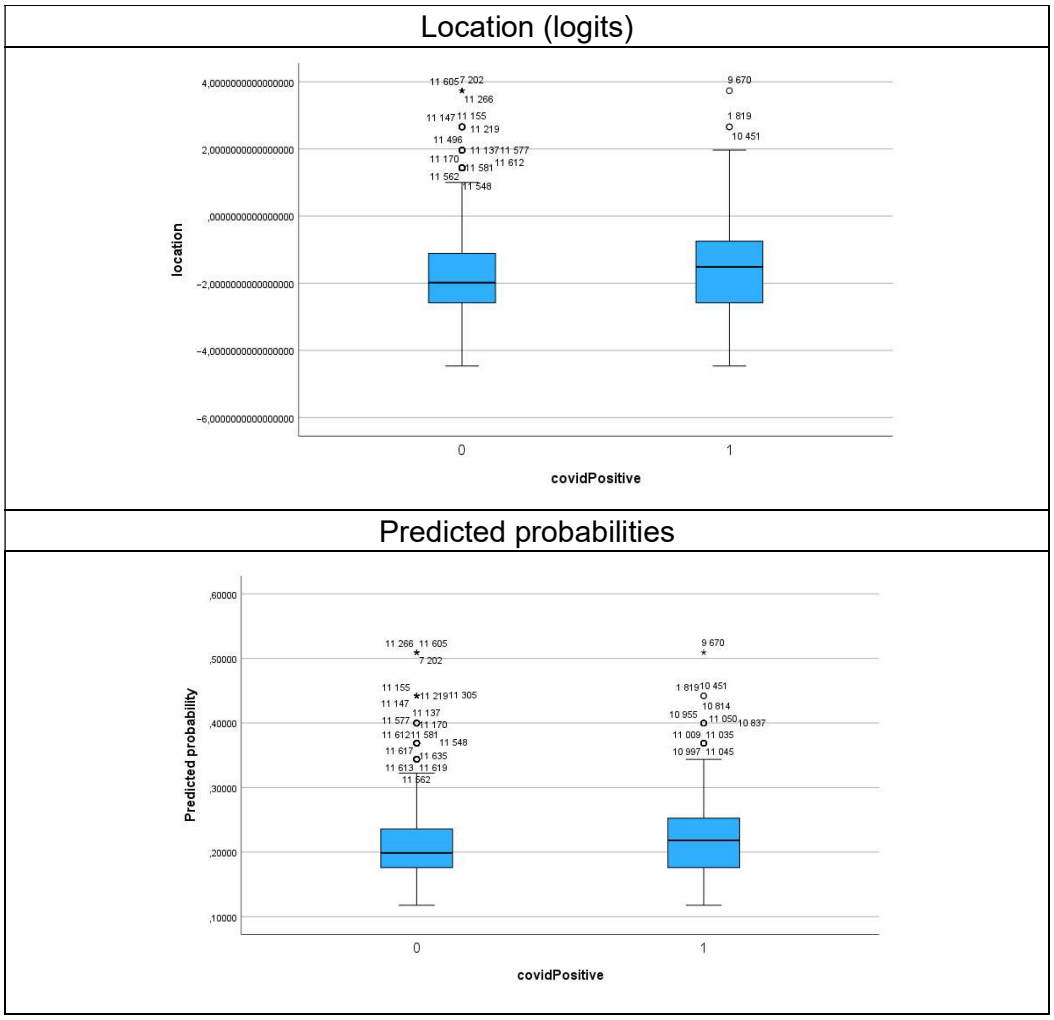


Figure 10. Box plots of logits (individual locations) from the Rasch analysis and probability of being COVID-positive, estimated from the logistic regression, for Model 3.

In Figure 11, the ROC curve is presented, with an AUC of 0.599 (95% CI 0.586; 0.611). The Precision-Recall Curve is also presented. Including sex and age in the logistic model only slightly improved the AUC.

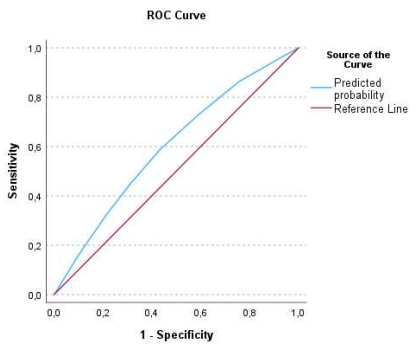
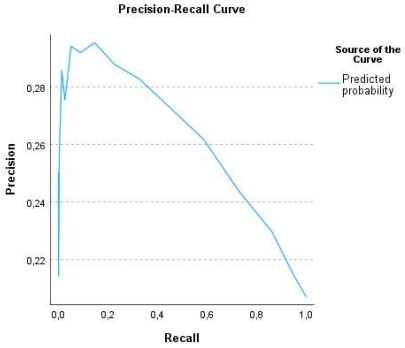
ROC curve	Precision-Recall Curve
	

Figure 11. Predicting positive or negative PCR test result. The ROC Curve and the precision-recall curve for Model 3.

Model 7

Predicting hospitalization

To predict hospitalization, only the individuals with a positive PCR test were included.

A graphical description of the location values and the prediction probabilities from the prediction model is shown in Figure 12.

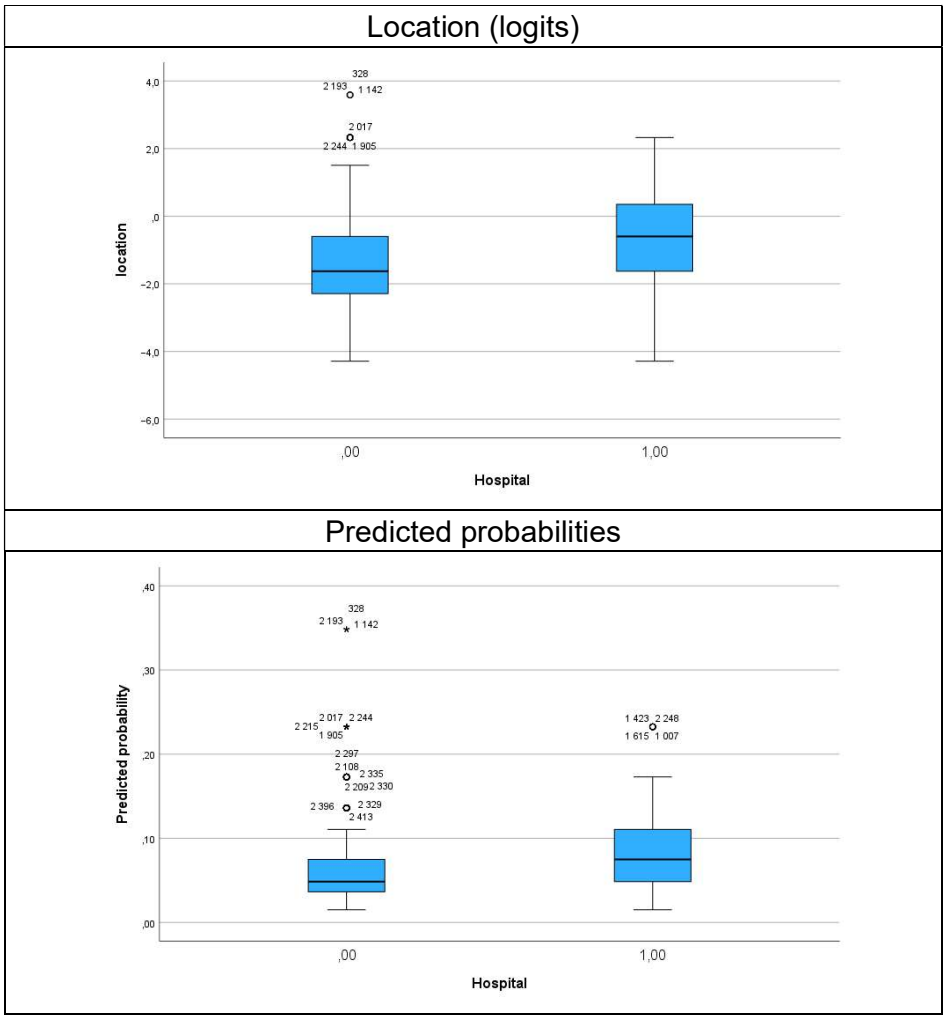


Figure 12. Model 7, box plots of logits (individual locations) from the Rasch analysis and probability of being COVID-positive, estimated from the logistic regression.

In Figure 13 the ROC curve is presented, with an AUC of 0.66 (95% CI 0.611; 0.707). The precision-recall curve is also presented. Including sex and age in the logistic model only slightly improved the AUC.

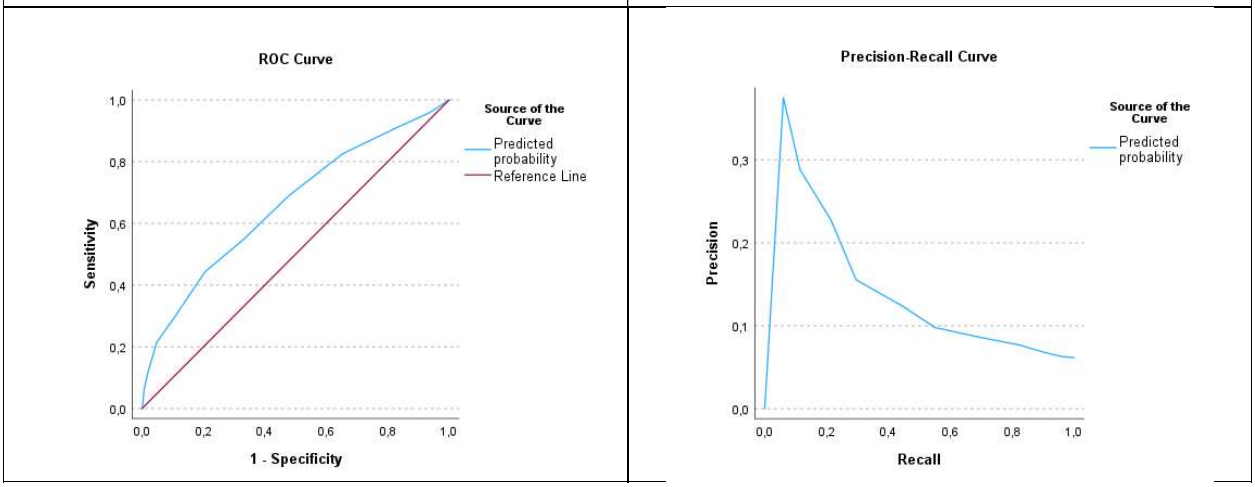
ROC curve	Precision-Recall Curve
	

Figure 13. Predicting hospitalization. The ROC curve and the precision-recall curve (positive prediction value-sensitivity) for Model 7.

Predicting positive or negative PCR test result

To predict a positive PCR test result or not, the full dataset of individuals with either a positive or a negative PCR test was used.

A graphical description of the location values and the prediction probabilities from the prediction model is shown in Figure 14.

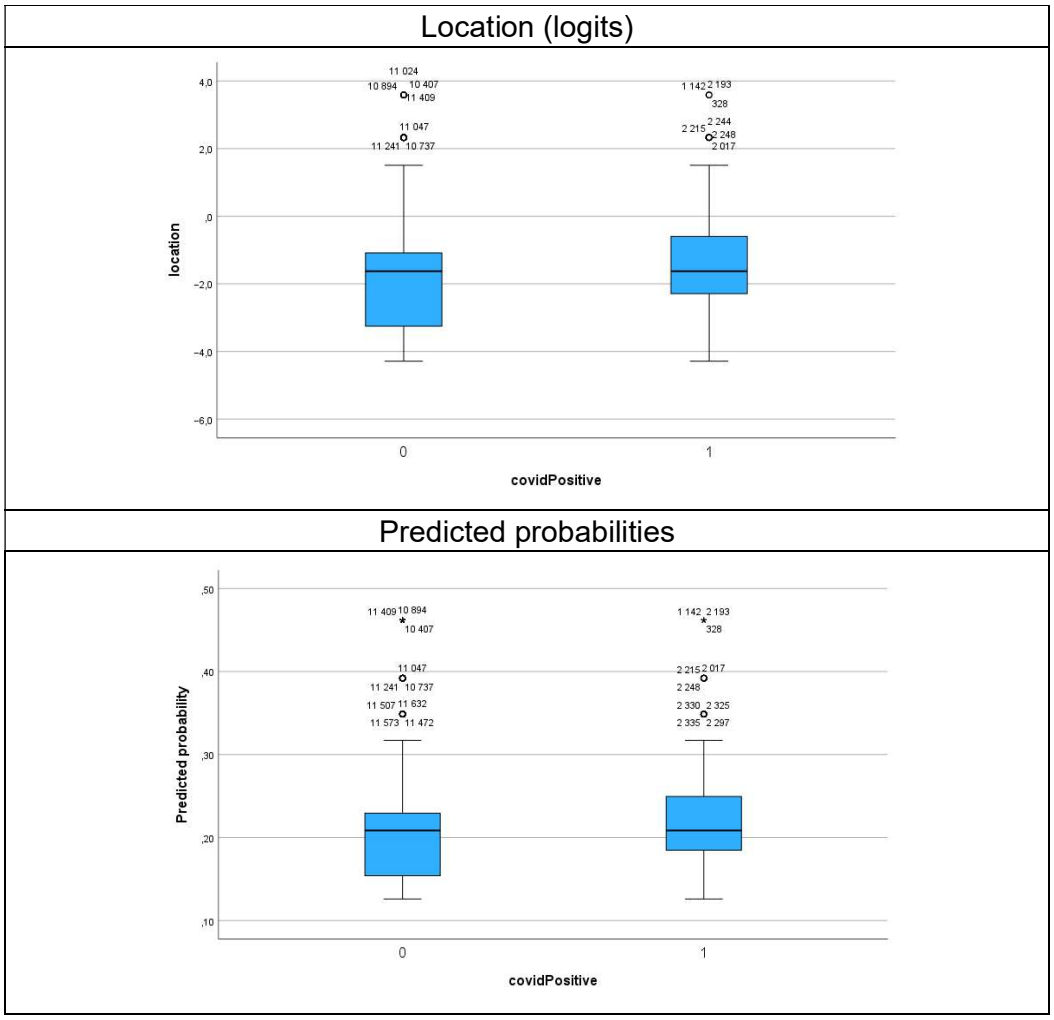


Figure 14. Model 7, box plots of logits (individual locations) from the Rasch analysis and probability of being COVID-positive, estimated from the logistic regression.

In Figure 15 the ROC curve is presented, with an AUC of 0.595 (95% CI 0.583; 0.608). The precision-recall curve is also presented. Including sex and age in the logistic model only slightly improved the AUC.

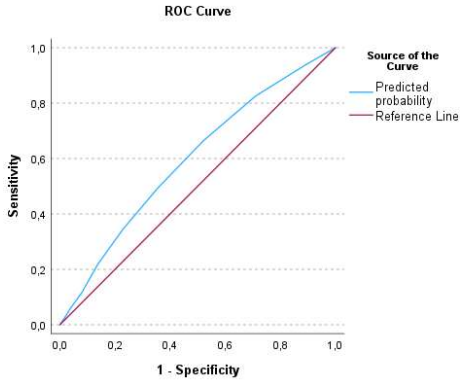
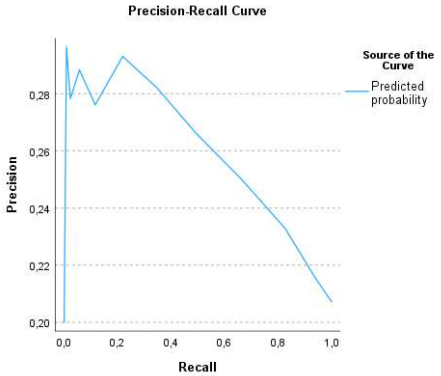
ROC curve	Precision-Recall Curve
	

Figure 15. Predicting positive or negative PCR test result. The ROC curve and the precision-recall curve for Model 7.

Discussion

Our aim was to investigate patterns of symptoms and their connection to COVID-19 infection and infection severity for the first 5 days after the start of symptoms.

Among individuals with positive PCR tests, the most frequently reported symptoms were headache (74%), fever (59%), sore throat (59%), and loss of smell (56%). In contrast, sore throat (78%) followed by headache (68%) were the most common symptoms among those with negative PCR tests. Other symptoms had a prevalence of 40% or lower, and loss of smell was reported by 15% of the individuals with negative PCR tests.

The fit to the Rasch model for the reported symptoms during the 5-day periods was found to be moderate. Low-scale values were associated with more common symptoms, such as headache and fever, whereas high-scale values represented uncommon symptoms like blisters on the feet. The predictive ability of the scale for hospitalization among individuals with positive PCR tests was also moderate, with an AUC of 0.66

Our findings are consistent with several other studies that have reported a high prevalence of similar symptoms in groups with COVID-19 infection (Bowyer et al., 2023; Tandan et al., 2021; Zens et al., 2020). However, the presence of many similar symptoms among individuals both with and without COVID-19 infection limits the utility of symptoms alone as a clear predictor of COVID-19 infection. It is essential to compare symptom patterns between COVID-19-positive and negative individuals to improve the screening process. Some studies have investigated associations between symptom patterns and variables identifying COVID-19 patients or those with severe COVID-19. However, the predictive capacity of these associations is often not evaluated.

As a result, the hypothesis that symptoms could form a scale indicating the likelihood of a COVID-19 infection was not fully supported by our study. The low-to-moderate fit to the Rasch model suggests that symptoms may not adhere to a ‘probability hierarchy’ and an alternative approach could be more useful, such as text clustering,

whereby symptoms are grouped based on their coexistence (Millar et al., 2022; Sudre et al., 2021).

Regarding the predictive ability of our scale, the AUC value of 0.6 for distinguishing between positive and negative PCR tests indicates limited discriminatory capacity. Remember that this is in the relation to screen for positive PCR tests, by using early symptoms (first 5-days of infection). Generally, ROC curves with an AUC≤0.75 are not considered clinically useful (Fan et al., 2006). Furthermore, the PSR in our Rasch analysis was 0.6, falling below the recommended 0.7 threshold required for effectively distinguishing between distinct groups, such as hospitalized and not hospitalized individuals. Only a few studies have examined the prediction capabilities of symptom patterns, and even fewer have explored their ability to screen for COVID-19 positives. In a study of real-time tracking of self-reported symptoms to predict potential COVID-19 a AUC of 0.76 was achived, with loss of smell and fatigue as key predictors (Menni et al., 2020), but here the full symptom period was used.

One study that predicted the risk of respiratory support requirement based on six clusters of symptom patterns showed promising results, with an AUC of 78.8 (Sudre et al., 2021).

The representation of COVID-19 cases in Swedish data during the first half of 2020 is limited because only more severe cases were tested, and thus, the data primarily reflects individuals with severe symptoms. Both COVID-positive and COVID-negative cases in this dataset are associated with severe symptoms, which suggests the possible prevalence of an ongoing severe disease.

One notable finding in the Swedish data is that the difference, between COVID-positive and COVID-negative cases, in age and hospitalization was smaller than initially hypothesized. This finding could be a contributing factor to the study's inability to establish a strong Rasch-based scale for assessing the presence or severity of COVID-19. It is worth noting that confirmed cases in the UK and US data may represent a more general population group compared to the Swedish data. In Sweden, in the beginning of the pandemic, testing of COVID-19 was not done in the general population, but mostly in seriously ill people who sought hospital contact.

In the present study, the sample sizes in the context of Rasch models were notably large, which can pose challenges for several tests assessing model and item fit (Müller, 2020). Nevertheless, the Andersen CLR test for model fit and the conditional infit/outfit tests, coupled with parametric bootstrap infit/outfit for item fit, have proven to be robust and suitable for large sample sizes. Because RUMM2030 is not designed to implement tests used for large sample sizes, eRm and iarm were used instead.

For sensitivity analysis and generating graphics, RUMM2030 was used as a complement to eRm and iarm. This approach allowed for a comprehensive assessment of the data, enabling a more accurate and insightful analysis.

In summary, symptom clustering holds significant promise in unraveling coexistence patterns of COVID-19 symptoms and has the potential to serve as a valuable screening tool for identifying severe cases. However, additional research using predictive models and comparative analyses is necessary to fully understand these patterns and their practical applications.

To establish more robust and accurate predictions, further studies should focus on developing predictive models and comparing COVID-19 positive and negative individuals. These efforts are essential in deepening our understanding of symptom patterns and their practical application in managing and mitigating the impact of COVID-19.

In conclusion, our findings indicate that predicting COVID-19 severity is feasible, making further research in this field imperative.

References

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Bowyer, R. C. E., Huggins, C., Toms, R., Shaw, R. J., Hou, B., Thompson, E. J., Kwong, A. S. F., Williams, D. M., Kibble, M., Ploubidis, G. B., Timpson, N. J., Sterne, J. A. C., Chaturvedi, N., Steves, C. J., Tilling, K., & Silverwood, R. J. (2023). Characterising patterns of COVID-19 and long COVID symptoms: evidence from nine UK longitudinal studies. *European Journal of Epidemiology*, *38*(2), 199–210. https://doi.org/10.1007/s10654-022-00962-6

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *CJEM*, *8*(01), 19–20. https://doi.org/10.1017/S1481803500013336

Kennedy, B., Fitipaldi, H., Hammar, U., Maziarz, M., Tsereteli, N., Oskolkov, N., Varotsis, G., Franks, C. A., Nguyen, D., Spiliopoulos, L., Adami, H.-O., Björk, J., Engblom, S., Fall, K., Grimby-Ekman, A., Litton, J.-E., Martinell, M., Oudin, A., Sjöström, T., … Fall, T. (2022). App-based COVID-19 syndromic surveillance and prediction of hospital admissions in COVID Symptom Study Sweden. *Nature Communications*, *13*(1), 2110. https://doi.org/10.1038/s41467-022-29608-7

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The **eRm** Package for the Application of IRT Models in *R*. *Journal of Statistical Software*, *20*(9). https://doi.org/10.18637/jss.v020.i09

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S.,

Chan, A. T., Steves, C. J., & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*, *26*(7), 1037–1040. https://doi.org/10.1038/s41591-020-0916-2

Millar, J. E., Neyton, L., Seth, S., Dunning, J., Merson, L., Murthy, S., Russell, C. D., Keating, S., Swets, M., Sudre, C. H., Spector, T. D., Ourselin, S., Steves, C. J., Wolf, J., Docherty, A. B., Harrison, E. M., Openshaw, P. J. M., Semple, M. G., Baillie, J. K., … Young, P. (2022). Distinct clinical symptom patterns in patients hospitalised with COVID-19 in an analysis of 59,011 patients in the ISARIC-4C study. *Scientific Reports*, *12*(1), 6843. https://doi.org/10.1038/s41598-022-08032-3

Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, *7*(1), 5. https://doi.org/10.1186/s40488-020-00108-7

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*(1), 1–18. https://doi.org/10.1348/014466506X96931

Sudre, C. H., Lee, K. A., Lochlainn, M. N., Varsavsky, T., Murray, B., Graham, M. S., Menni, C., Modat, M., Bowyer, R. C. E., Nguyen, L. H., Drew, D. A., Joshi, A. D., Ma, W., Guo, C.-G., Lo, C.-H., Ganesh, S., Buwe, A., Pujol, J. C., du Cadet, J. L., … Ourselin, S. (2021). Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. *Science Advances*, *7*(12). https://doi.org/10.1126/sciadv.abd4177

Tandan, M., Acharya, Y., Pokharel, S., & Timilsina, M. (2021). Discovering symptom patterns of COVID-19 patients using association rule mining. *Computers in Biology and Medicine*, *131*, 104249. https://doi.org/10.1016/j.compbiomed.2021.104249

World Health Organization. (2020). *Coronavirus disease 2019 (COVID-19) situation report–51. Geneva, Switzerland: World Health Organization*.

Https://Www.Who.Int/Docs/Default-Source/Coronaviruse/Situation-Reports/20200311-Sitrep-51-Covid-19.Pdf?Sfvrsn=1ba62e57_10.

Zens, M., Brammertz, A., Herpich, J., Südkamp, N., & Hinterseer, M. (2020). App-Based Tracking of Self-Reported COVID-19 Symptoms: Analysis of Questionnaire Data. *Journal of Medical Internet Research*, *22*(9), e21956. https://doi.org/10.2196/21956