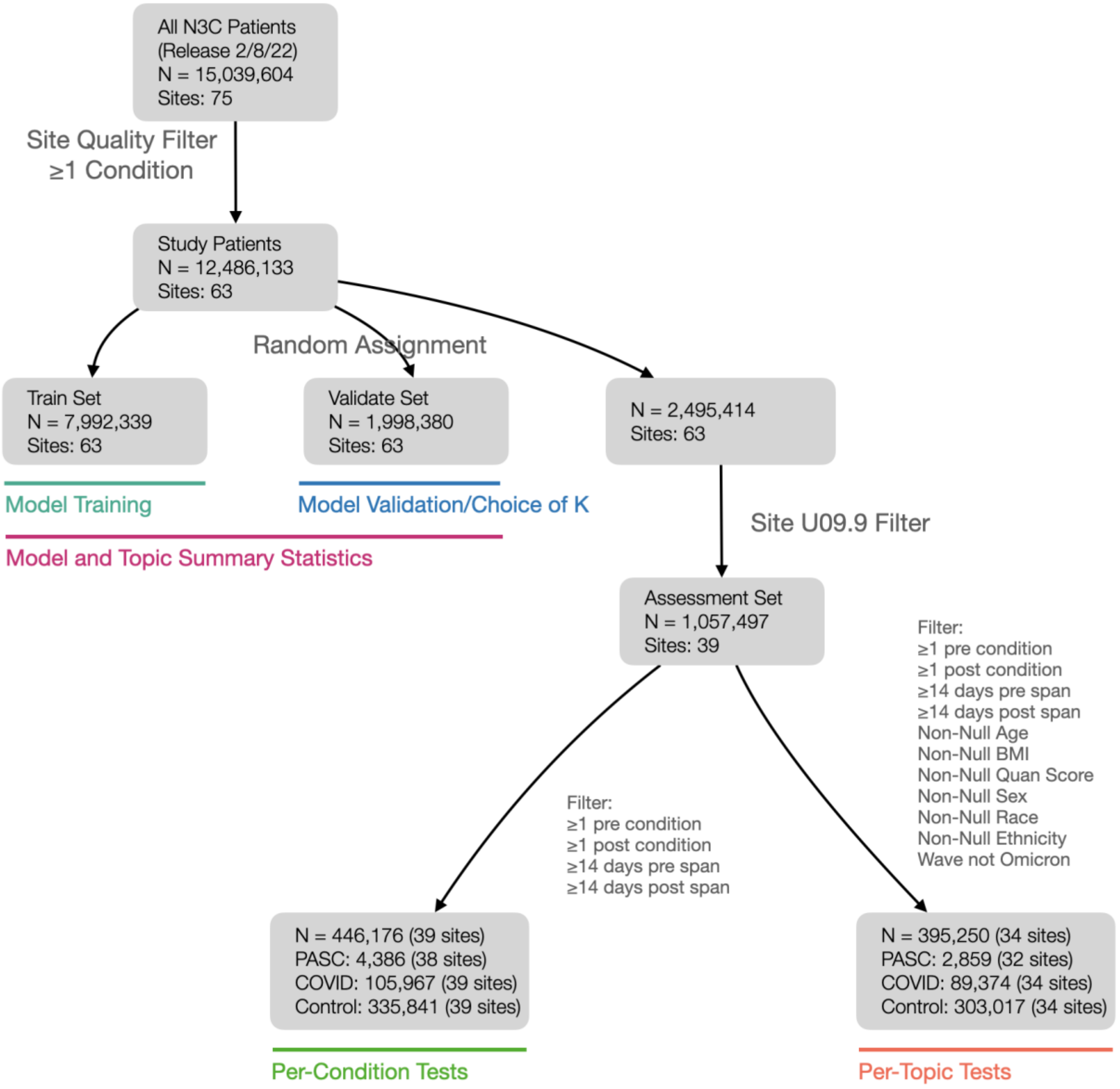
Supplemental Figures

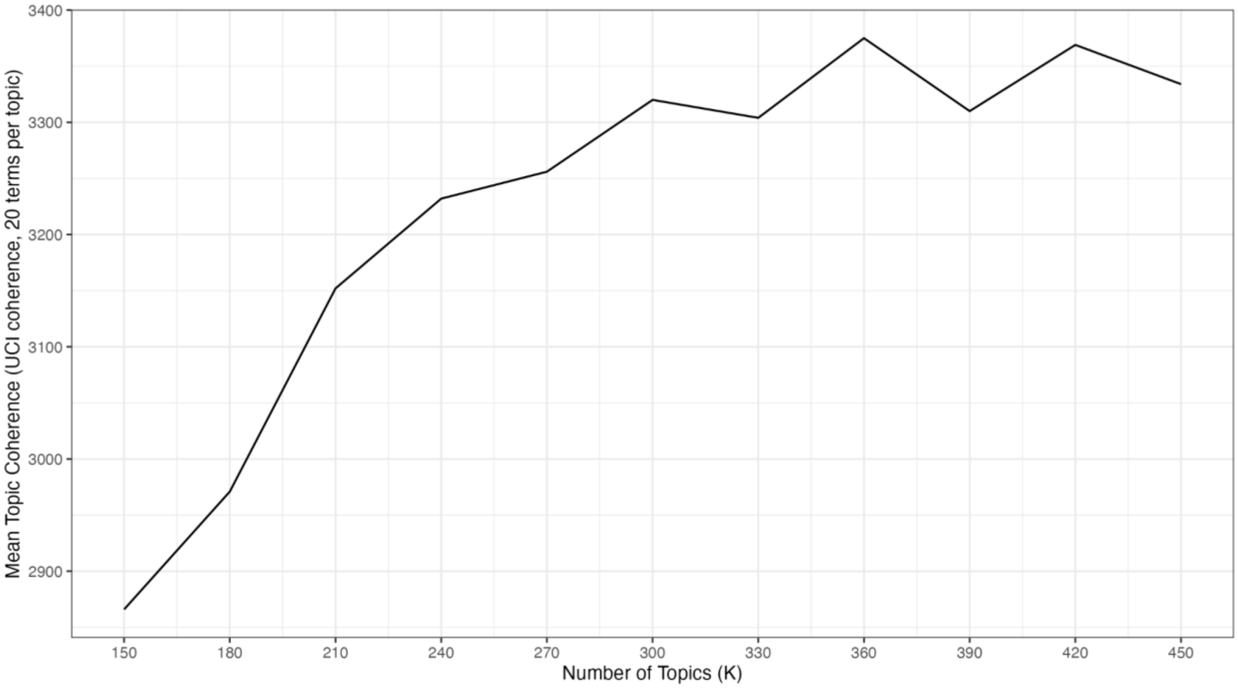
Suppl. Figure 1

Consort diagram illustrating stratification of patients into sets and cohorts, number of unique sites represented by those groups, and how each is used in analysis. The site quality filter removed sites with inpatient serum creatinine or white blood cell count results for fewer than 25% of patients, the site U09.9 filter removed patients from sites with no U09.9 diagnoses, and filter variables are as described for specific tests (see Suppl. Methods).



Suppl. Figure 2

Mean topic coherence scores for LDA models varying the number of topics generated (K). Topic coherences are computed as intrinsic UCI Coherence[30] using the top 20 terms per topic. UCI coherence evaluates, for all term pairs amongst these top 20, how frequently they occur together in patient histories compared to the expectation assuming terms occur independently, on the validation data set. K=300 was chosen as the final number of topics.



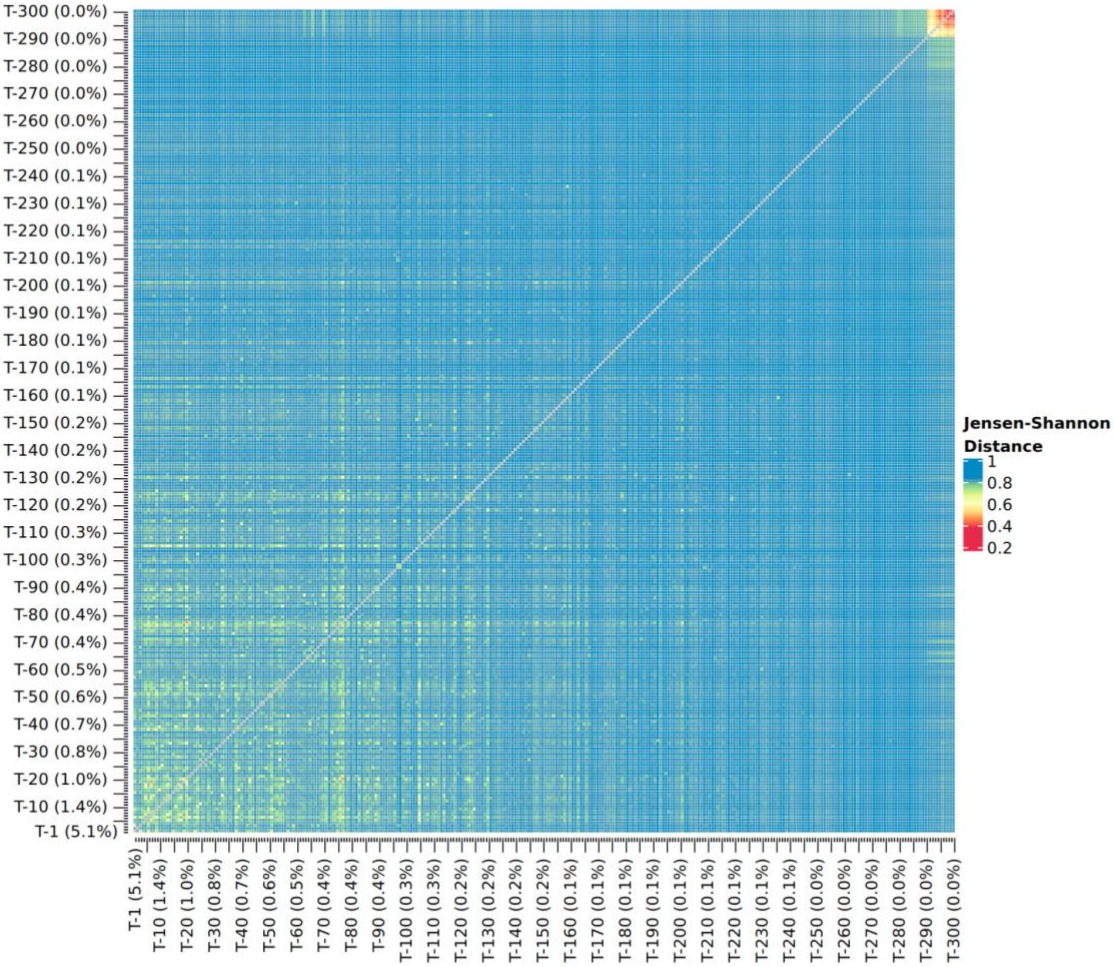
Suppl. Figure 3

Full topic clouds for all 300 topics generated and visualizations of corresponding contrasts.

Available at https://doi.org/10.5281/zenodo.11188766.

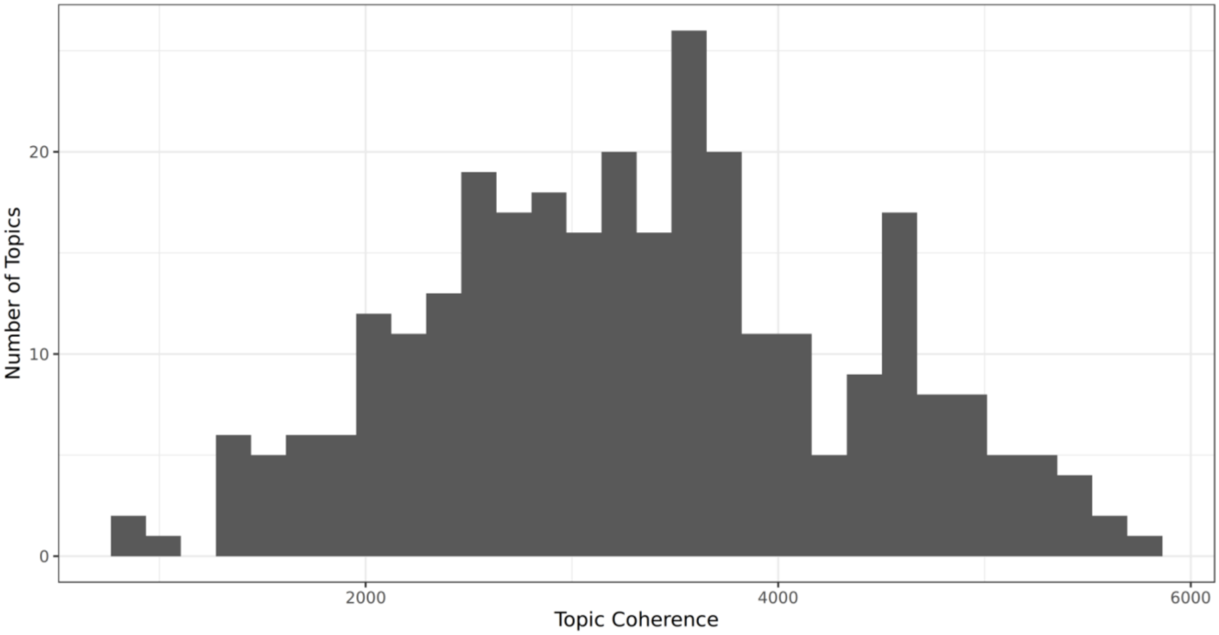
Suppl. Figure 4

Topic/topic dissimilarity as Jensen-Shannon Distance. Topic self-distances of 0 are not shown.



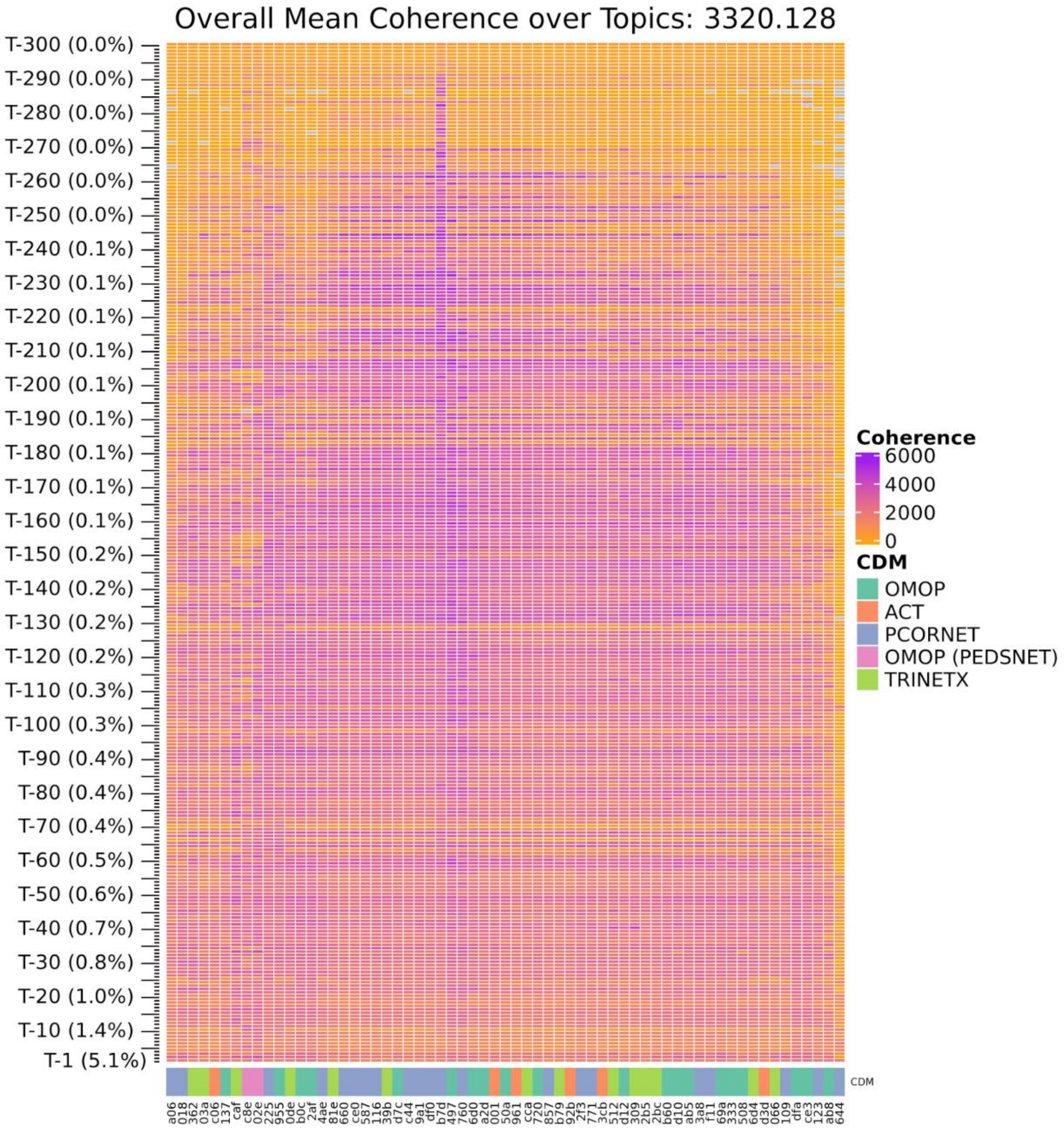
Suppl. Figure 5

Histogram of topic coherence values.



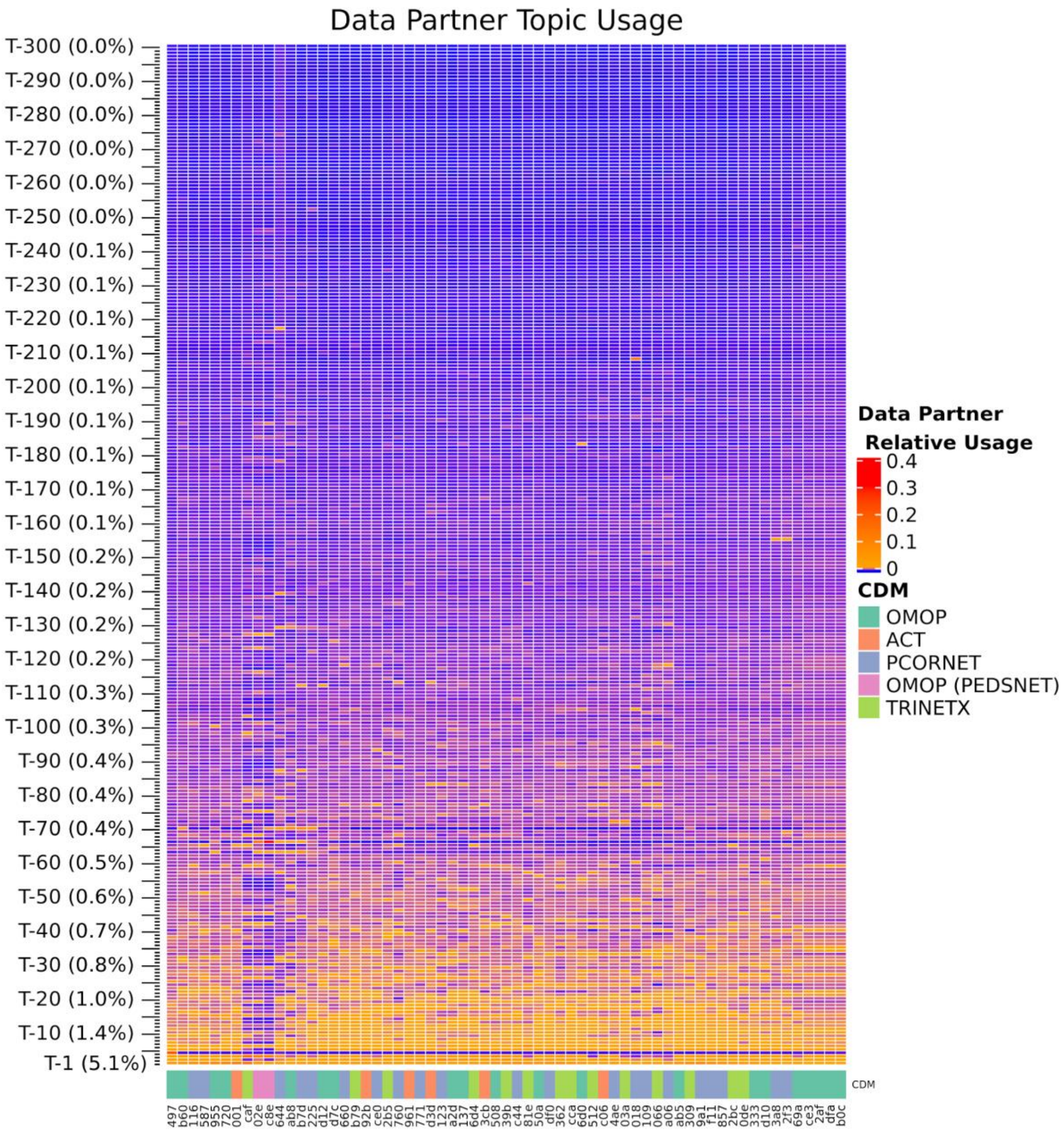
Suppl. Figure 6

Mean UCI coherence scores per topic and contributing data site (ID anonymized). Site identifiers are masked, but labeled with the source common data model in use at the site.



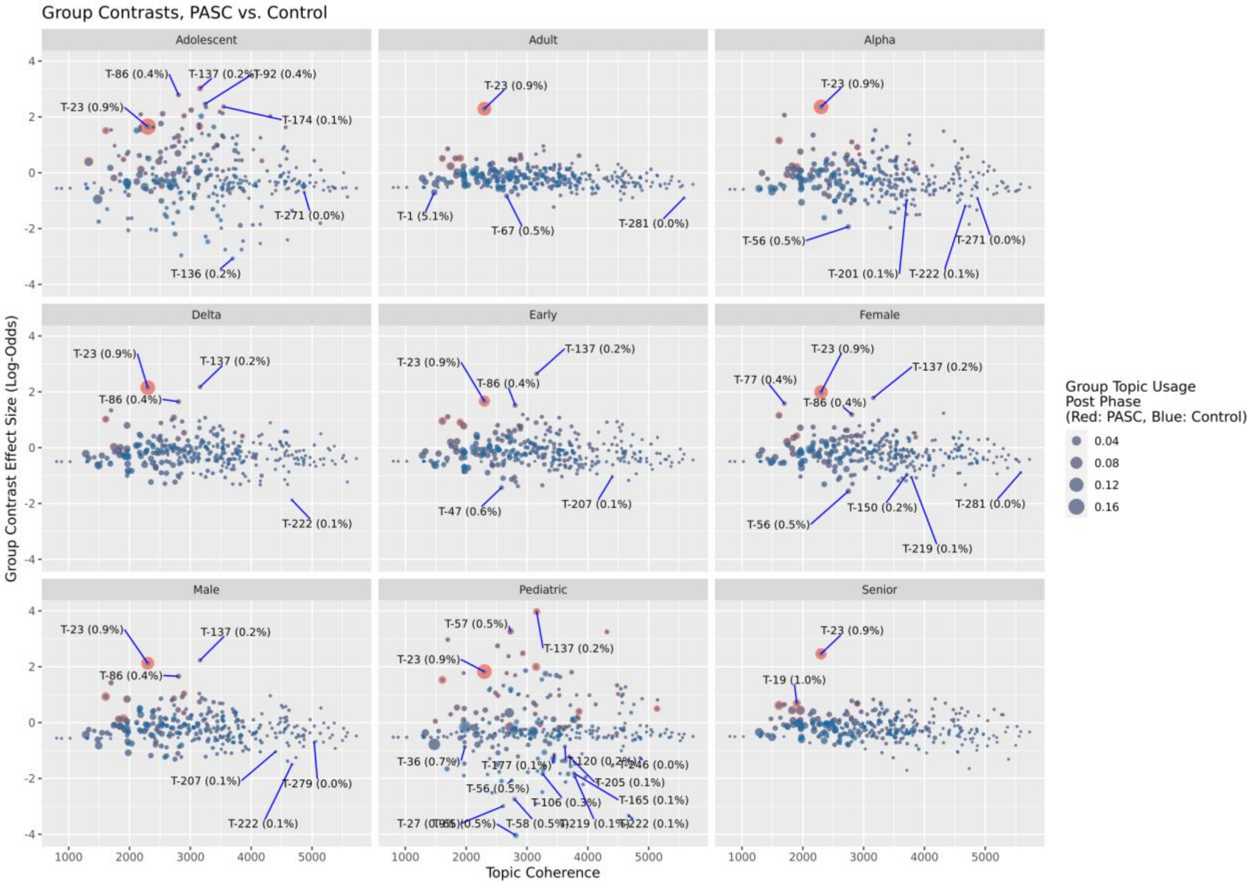
Suppl. Figure 7

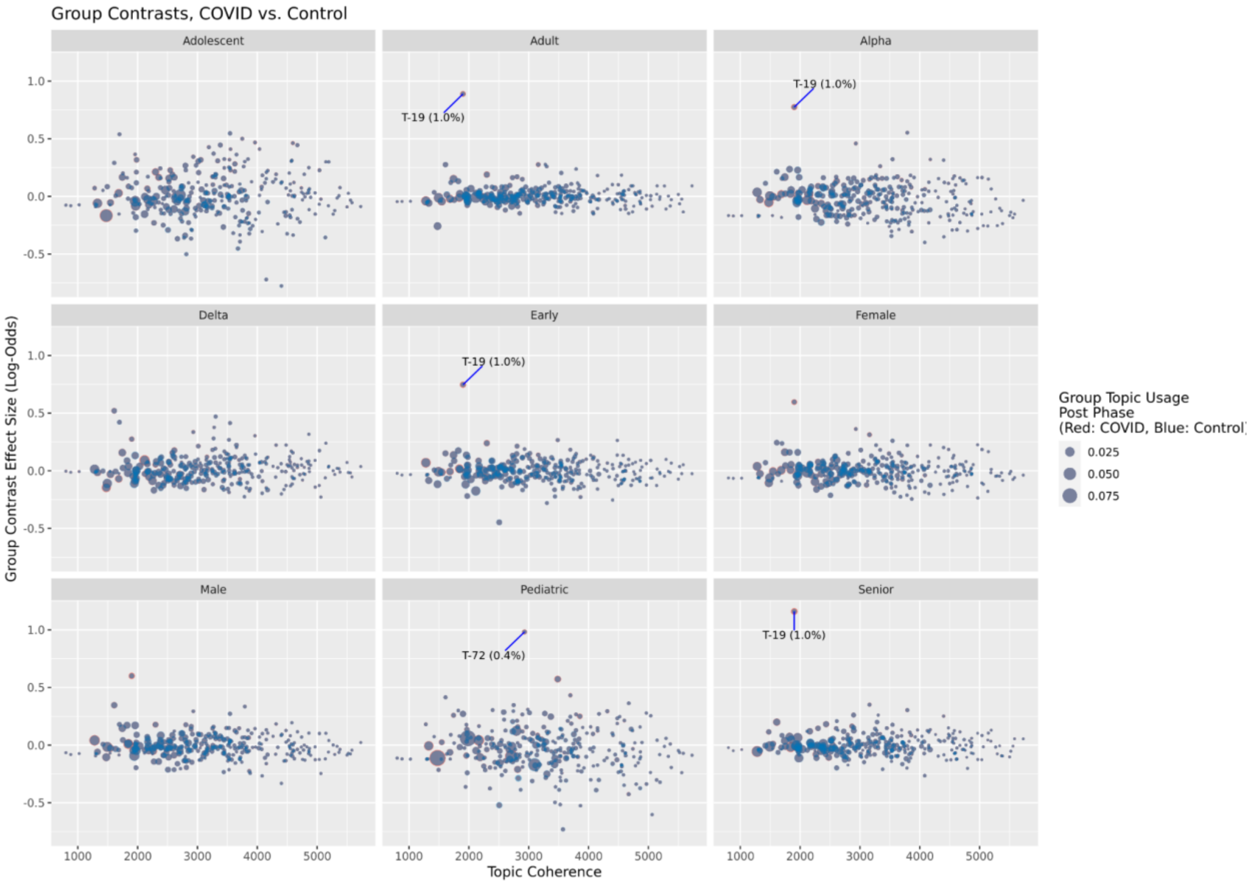
Relative usage of topics per contributing site (ID anonymized). For a given site and topic, relative usage is computed as the sum of assigned weights to that topic for patients from that site divided by the number of patients, representing a distribution over topics per site.



Suppl. Figure 8

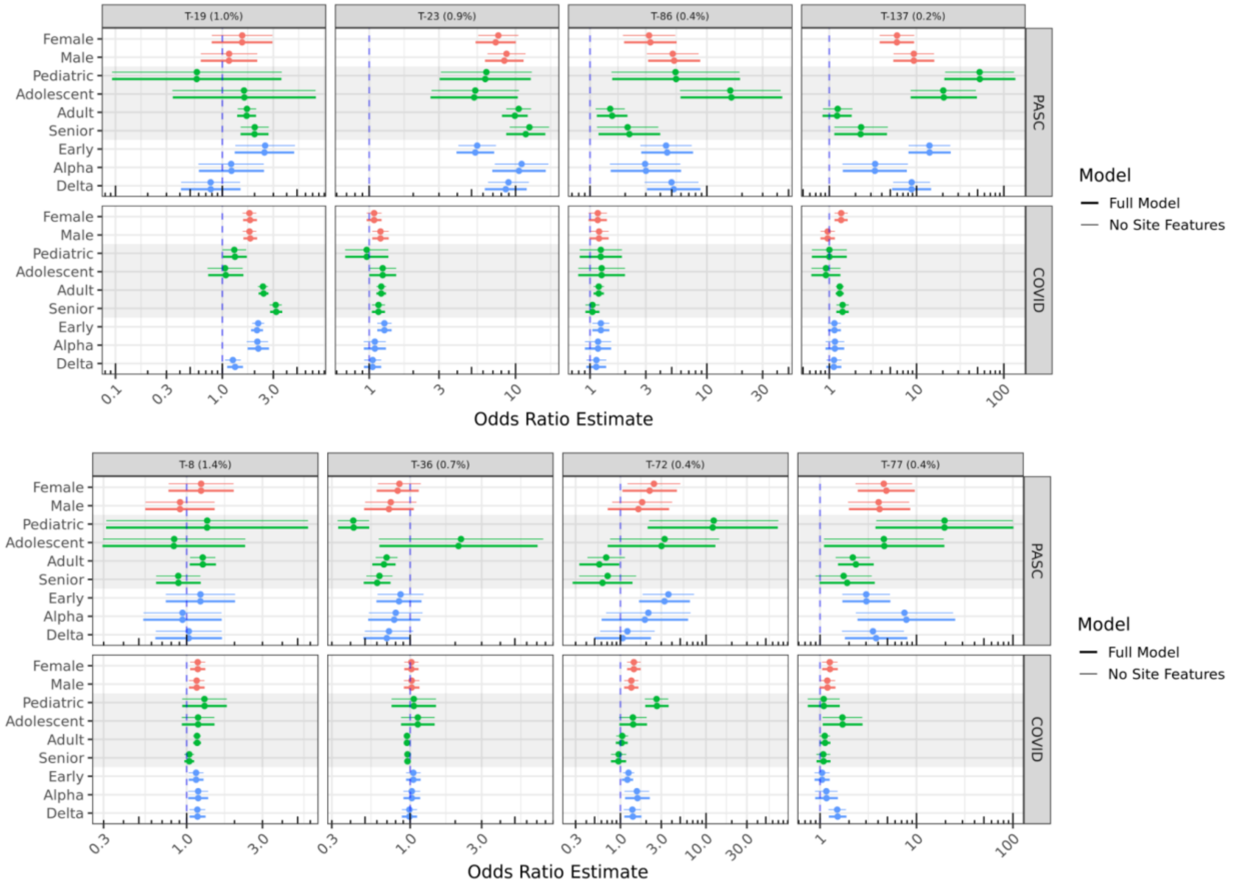
Per-topic coherence (horizontal axis) vs. contrast effect sizes (log-odds scale, vertical axis) for tested groups (panels) in PASC vs. Control (top) and COVID vs. Control (bottom) contrasts. Labeled topics are those with statistically significant log-odds differences of >1 or <-1 (OR >2 or <0.5). Points are sized and colored according to mean topic usage for the group and cohort in the post-infection phase, with blue points representing Control patients and red points representing PASC (top) or COVID (bottom) patients.




Group Contrasts, COVID vs. Control
Adolescent
Adult
Alpha
Delta
Early
Female
Male
Pediatric
Senior
T-19 (1.0%)
T-19 (1.0%)
T-19 (1.0%)
T-72 (0.4%)
T-19 (1.0%)
Group Contrast Effect Size (Log-Odds)
1.0
0.5
0.0
-0.5
Topic Coherence
1000
2000
3000
4000
5000
Group Topic Usage
Post Phase
(Red: COVID, Blue: Control)
0.025
0.050
0.075

Suppl. Figure 9

Results for Figures 4 (top) and 5 (bottom) for models with and without site-level covariates of topic usage, percentage of PASC patients, and source common data model.



Supplemental Methods

Minimal Site Quality Filters

EHR data from the National COVID Cohort Collaborative (N3C), released Aug. 2, 2022 represent records from 75 contributing sites. All analyses were restricted to data from 63 sites passing minimal quality checks: sites were excluded if greater than 25% of inpatient visits were not accompanied by serum creatinine or white blood cell count measures (N=11), or if greater than 5% of COVID-19 confirmed patients were indicated as inpatient continuously for 200 or more days prior to and including their confirmed COVID-19 date (as potential long-term care facilities, N=1).

Model Training

Model training utilized the online Latent Dirichlet Allocation (LDA) method of Hoffman et al.[24] as implemented in Apache Spark (pyspark.ml.clustering.LDA) version 3.2.1.[29] Parameters used include k (the number of topics, 300 in the final model), seed (42, a random seed to initialize the training), and maxIter (200, providing 10 passes over the training data in batches of 5% each). Determination of condition-topic and topic-patient distributions were produced by the fitted LDA model.

Topic Annotations

Each topic is annotated with three values: U, representing the relative usage of the topic by total weight assigned to patients (range 0-100%), H, a measure how uniformly the topic is used by N3C-contributing sites (range 0-1, with values closer to 0 being site-specific), and C, a measure of each topics' coherence compared to the mean over all topics. All three are computed over the training and validation sets.

U is computed as the sum over patients of the weight assigned to the topic, divided by the number of patients (which is also the total weight assigned over all topics).

H is computed as the information entropy of the relative usage of the topic across sites, normalized to a maximum value of 1.0 when the usage is uniformly distributed. Relative usage for a given site is computed as the total weight assigned to the topic for patients from the site, divided by the total number of patients from that site.

Per-topic coherence C is calculated for each topic using the UCI Coherence metric (see Model Validation below). These values are not meant to be interpreted on an absolute scale, but since they are normally distributed amongst topics (Suppl. Figure 4) we adjust them to z-scores for comparative use.

Jensen-Shannon Distance

Jensen-Shannon Distance between topics t_i and t_j is a true metric and is defined as the square root of the Jensen-Shannon divergence:

$$\text{JS Distance}(t_i, t_j) = \sqrt{\frac{\sum_{c \in terms} c_{t_i} \log\left(c_{t_i}/M\right)}{2} + \frac{\sum_{c \in terms} c_{t_i} \log\left(c_{t_i}/M\right)}{2}}\,,$$

where $c_{t_x} = p(c|t_x)$ (the probability assigned to term c in topic t_x) and M is $(c_{t_i} + c_{t_j})/2$.

Topic Term Relevance

Term relevance provides a measure of term-topic-specificity, with values greater than zero indicating terms more likely for the topic than overall.[33] For term c_i and topic t_j, we define relevance as

$$\text{relevance}(c_i) = \ln \frac{p(c_i|t_j)}{p(c_i)}\ .$$

Model Validation

UCI coherence for a given topic $t_i$is computed over the top N terms by probability for the topic, where we used N = 20. Letting T_i be the set of top 20 terms for t_i, a sum score is computed for each distinct pair of terms a and b, where the score for a given pair is the log of the measured probability of their occurring together in a patient compared to the joint probability assuming independence. To avoid undefined scores, 0 is used for pairs where the denominator is 0, and 1 is added to the joint probability.[30]

$$\text{Coherence}(t_i) = \sum_{b,c \in T_i,\, b<c} \begin{cases} \log_2\left(\frac{1+p(b \cap c|t_i)}{p(b|t_i) \cdot p(c|t_i)}\right) & \text{if } p(b \mid t_i) \cdot p(c \mid t_i) > 0 \\ 0 & \text{else} \end{cases}$$

Overall model quality was evaluated as the mean of coherence scores across topics, computed over the validation dataset only.

Per-Condition Tests

All tests were performed in R v3.5.1.[61] As described in the main text, patients in the test data set were included for evaluation of new-onset conditions if they satisfied requirements for being in the PASC, COVID, or Control cohorts. The top 20 conditions from each topic with relevance score > 0 were evaluated by considering only patients without the condition in the pre phase, comparing counts of PASC (and COVID) patients later indicated and not indicated for the post phase, to those same counts in the Control cohort. R's fisher.test() was used with simulate.p.value = TRUE to support tests where counts are large.[34] Reported p values were multiple-test corrected using Bonferroni's method.

BMI and Quan Comorbidity Scores

Patient BMI values used in modeling were the maximum over those reported after Jan. 1 2018, or the maximum of those computed as weight/(height^2) if no BMI measurement was directly available. Weight values outside 5kg–300kg and height values outside 0.6m–2.43m were excluded from BMI calculations. Quan comorbidity scores[40] were computed from available source ICD code prefixes as shown in Suppl. Table 7.

Topic Regression Tests

Regression models were fitted using `geepack` v1.3.9,[38] with contrasts computed using `emmeans` v1.8.9.[62] Individual patient histories defined by their pre- and post- phase data were assigned topic probability distributions by the fitted LDA model. For each topic, we fitted a logistic regression model with outcome variable being the model-assigned topic probability as the trial success rate with equal weight, from covariates phase (pre or post), cohort (PASC, COVID, or Control), patient life stage and wave of the index date (see main Methods), sex, race, Quan comorbidity score, BMI, source CDM (PCORnet, ACT, OMOP, TrinetX, and OMOP (PedsNet)). To account for potential differential usage of PASC labels or topics, we also included percentage of patients at the given patients' site in the PASC cohort, and usage of the topic by the patients' site relative to all sites (summing to 1.0 across sites). Interactions were included for terms of interest for contrasts using the R/geepack formula `topic_probability ~ phase * cohort * (index_wave + sex + life_stage) + site_percent_pasc * phase * cohort + site_relative_topic_usage + race + quan_score + bmi + cdm`. Only patients from the assessment set with complete information for all variables were included.

Supplemental Tables

Suppl. Table 1

OMOP Concepts excluded from model training, evaluation, and testing.

Concept Name	OMOP Concept Id
No matching concept	0
Clinical finding	441840
COVID-19	37311061
Viral disease	440029
Disease due to coronaviridae	4100065
Sexually abstinent	764423
Single current sexual partner	4043045
New sexual partner	44813701
Sexually active with men	43021202
Single historical sexual partner	43021216
Number of current sexual partners - finding	4276728
Bigamy	4336540
Sexual activity - two to three times per month	4012347
Sexual activity - two to three times per week	4012202
Finding of number of historical sexual partners	43021214
No longer sexually active	4043041
Multiple current sexual partners	4038723
Sexually active with transgender person	43021204
Number of sexual partners - finding	4269990
Satisfactory sexual experience	44811373
Sexual activity - daily	4012377
Currently not sexually active	4012376
Never been sexually active	4145811
Fornication	4031991
Sexual activity - monthly	4012348
Sexual activity - weekly	4012203
Sexual contact with high risk partner	44789379
Finding of frequency of sexual activity	4188013
Engages in sexual activity outside marriage	43021163
Sexually active with women	43021203
Purposely unmarried and sexually abstinent	43021238
Sex within a relationship only	4021660
Sexually active in last month	37017764
Sexually active	4043042
Finding relating to sexual activity	4114865
Sexually active in last year	37017763
Engages in sexual activity before marriage	43021162
Sexually active in last six months	37017762
Multiple historical sexual partners	43021215

Suppl. Table 2

OMOP Concepts describing COVID-19 PCR or Antigen tests.

Concept Name	OMOP Concept Id
SARS-CoV-2 (COVID-19) N gene [Presence] in Respiratory specimen by Nucleic acid amplification using CDC primer-probe set N2	586525
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Saliva (oral fluid) by NAA with probe detection	36032174
SARS-related coronavirus RNA [Presence] in Specimen by NAA with probe detection	723472
SARS-CoV-2 (COVID-19) N gene [Cycle Threshold #] in Specimen by Nucleic acid amplification using CDC primer-probe set N2	706155
SARS-CoV-2 (COVID-19) S gene [Cycle Threshold #] in Specimen by NAA with probe detection	723468
SARS-CoV-2 (COVID-19) N gene [#/volume] (viral load) in Respiratory specimen by NAA with probe detection	36661370
SARS-CoV-2 (COVID-19) S gene [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	723467
SARS-CoV-2 (COVID-19) N gene [Presence] in Serum or Plasma by NAA with probe detection	586520
SARS-CoV-2 (COVID-19) S gene [Presence] in Respiratory specimen by NAA with probe detection	723465
SARS-CoV-2 (COVID-19) [Presence] in Specimen by Organism specific culture	586516
SARS-CoV-2 (COVID-19) N gene [Cycle Threshold #] in Specimen by NAA with probe detection	706167
SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay	723477
SARS-CoV-2 (COVID-19) RNA [Log #/volume] (viral load) in Specimen by NAA with probe detection	715262
SARS-related coronavirus N gene [Cycle Threshold #] in Specimen by Nucleic acid amplification using CDC primer-probe set N3	706172
SARS-CoV-2 (COVID-19) RNA [Presence] in Saliva (oral fluid) by NAA with probe detection	715260
SARS-CoV-2 (COVID-19) S gene [Presence] in Serum or Plasma by NAA with probe detection	586519
SARS-CoV-2 (COVID-19) ORF1ab region [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	723469
SARS-CoV-2 (COVID-19) RNA [Cycle Threshold #] in Specimen by NAA with probe detection	586529
SARS-related coronavirus E gene [Presence] in Respiratory specimen by NAA with probe detection	586523
SARS-CoV-2 (COVID-19) ORF1ab region [Presence] in Saliva (oral fluid) by NAA with probe detection	36031506

Concept Name	OMOP Concept Id
SARS-CoV-2 (COVID-19) S gene [Presence] in Specimen by NAA with probe detection	723466
SARS-CoV-2 (COVID-19) RNA [Presence] in Nasopharynx by NAA with non-probe detection	723476
SARS-CoV-2 (COVID-19) N gene [Presence] in Saliva (oral fluid) by Nucleic acid amplification using CDC primer-probe set N1	36032258
SARS-CoV-2 (COVID-19) RNA [Presence] in Nasopharynx by NAA with probe detection	586526
SARS-related coronavirus E gene [Presence] in Serum or Plasma by NAA with probe detection	586518
SARS-CoV-2 (COVID-19) S gene [Presence] in Respiratory specimen by Sequencing	36031213
SARS-CoV-2 (COVID-19) RNA [Presence] in Nose by NAA with probe detection	757677
SARS-CoV-2 (COVID-19) N gene [Presence] in Specimen by Nucleic acid amplification using CDC primer-probe set N2	706154
SARS-CoV-2 (COVID-19) RNA panel - Respiratory specimen by NAA with probe detection	706158
SARS-CoV-2 (COVID-19) N gene [Presence] in Respiratory specimen by NAA with probe detection	706161
SARS-CoV-2 (COVID-19) RdRp gene [Cycle Threshold #] in Specimen by NAA with probe detection	723470
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Lower respiratory specimen by NAA with probe detection	36031652
SARS-CoV-2 (COVID-19) N gene [Presence] in Saliva (oral fluid) by NAA with probe detection	36661378
SARS-related coronavirus+MERS coronavirus RNA [Presence] in Respiratory specimen by NAA with probe detection	706159
SARS-related coronavirus E gene [Presence] in Specimen by NAA with probe detection	706174
SARS-CoV-2 (COVID-19) N gene [Presence] in Specimen by Nucleic acid amplification using CDC primer-probe set N1	706156
SARS-CoV-2 (COVID-19) RNA [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	586528
Measurement of Severe acute respiratory syndrome coronavirus 2 antigen	37310257
SARS-related coronavirus E gene [Cycle Threshold #] in Specimen by NAA with probe detection	706166
SARS-CoV-2 (COVID-19) Ag [Presence] in Upper respiratory specimen by Immunoassay	36032419
SARS-CoV-2 (COVID-19) RNA panel - Specimen by NAA with probe detection	706169
SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by NAA with non-probe detection	36031238

Concept Name	OMOP Concept Id
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Respiratory specimen by NAA with probe detection	706160
SARS-CoV-2 (COVID-19) N gene [Presence] in Nasopharynx by NAA with probe detection	715272
SARS-CoV-2 (COVID-19) N gene [Presence] in Nose by NAA with probe detection	757678
SARS-CoV-2 (COVID-19) RNA [Presence] in Saliva (oral fluid) by Sequencing	715261
SARS-CoV-2 (COVID-19) RNA [Presence] in Specimen by NAA with probe detection	706170
SARS-CoV-2 (COVID-19) N gene [Cycle Threshold #] in Specimen by Nucleic acid amplification using CDC primer-probe set N1	706157
SARS-CoV-2 (COVID-19) ORF1ab region [Presence] in Respiratory specimen by NAA with probe detection	723478
SARS-related coronavirus N gene [Presence] in Specimen by Nucleic acid amplification using CDC primer-probe set N3	706171
SARS-CoV+SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay	757685
SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by Sequencing	36661377
SARS-CoV-2 (COVID-19) N gene [Log #/volume] (viral load) in Respiratory specimen by NAA with probe detection	36661371
SARS-CoV-2 (COVID-19) RdRp gene [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	723471
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Upper respiratory specimen by NAA with probe detection	36031453
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Specimen by NAA with probe detection	706173
SARS-CoV-2 (COVID-19) N gene [Presence] in Specimen by NAA with probe detection	706175
SARS-CoV-2 (COVID-19) ORF1ab region [Cycle Threshold #] in Specimen by NAA with probe detection	706168
SARS-CoV-2 (COVID-19) N gene [Presence] in Respiratory specimen by Nucleic acid amplification using CDC primer-probe set N1	586524
SARS-CoV-2 (COVID-19) ORF1ab region [Presence] in Specimen by NAA with probe detection	723464
SARS-related coronavirus RNA [Presence] in Respiratory specimen by NAA with probe detection	706165
SARS-CoV-2 (COVID-19) RNA panel - Saliva (oral fluid) by NAA with probe detection	36032061
SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by NAA with probe detection	706163
SARS-CoV-2 (COVID-19) specific TCRB gene rearrangements [Presence] in Blood by Sequencing	36031944

Concept Name	**OMOP Concept Id**
SARS-CoV-2 (COVID-19) RNA [Presence] in Serum or Plasma by NAA with probe detection	723463

Suppl. Table 3

All indicators of COVID-19 infection (except for PCR and Antigen tests, Suppl. Table 3).

Concept Name	Concept Id
SARS-CoV-2 (COVID-19) IgG Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	706181
SARS-CoV-2 (COVID-19) IgA Ab [Units/volume] in Serum or Plasma by Immunoassay	723459
SARS-CoV-2 (COVID-19) IgM Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	706180
SARS-CoV-2 (COVID-19) IgM Ab [Presence] in DBS by Immunoassay	36659631
SARS-CoV-2 (COVID-19) IgM Ab [Titer] in Serum or Plasma by Immunofluorescence	36661373
SARS-CoV-2 (COVID-19) neutralizing antibody [Presence] in Serum by pVNT	757680
SARS-CoV-2 (COVID-19) IgG+IgM Ab [Presence] in Serum or Plasma by Immunoassay	723479
SARS-CoV-2 (COVID-19) Ab panel - Serum, Plasma or Blood by Rapid immunoassay	706176
SARS-CoV-2 (COVID-19) IgG Ab [Titer] in Serum or Plasma by Immunofluorescence	36661374
SARS-CoV-2 (COVID-19) IgM Ab [Units/volume] in Serum or Plasma by Immunoassay	706178
SARS-CoV-2 (COVID-19) IgA Ab [Presence] in Serum or Plasma by Immunoassay	723473
SARS-CoV-2 (COVID-19) neutralizing antibody [Titer] in Serum by pVNT	757679
SARS-CoV-2 (COVID-19) Ab [Presence] in Serum or Plasma by Immunoassay	586515
SARS-CoV-2 (COVID-19) IgG Ab [Units/volume] in Serum or Plasma by Immunoassay	706177
SARS-CoV-2 (COVID-19) S protein RBD neutralizing antibody [Presence] in Serum or Plasma by sVNT	36031734
SARS-CoV-2 (COVID-19) IgA Ab [Titer] in Serum or Plasma by Immunofluorescence	36661372
SARS-CoV-2 (COVID-19) Ab [Units/volume] in Serum or Plasma by Immunoassay	586522
SARS-CoV-2 (COVID-19) IgA+IgM [Presence] in Serum or Plasma by Immunoassay	757686
Measurement of Severe acute respiratory syndrome coronavirus 2 antibody	37310258
SARS-CoV-2 (COVID-19) IgG Ab [Presence] in Serum or Plasma by Immunoassay	723474
SARS-CoV-2 (COVID-19) Ab panel - Serum or Plasma by Immunoassay	706179
SARS-CoV-2 stimulated gamma interferon [Presence] in Blood	36031969
SARS-CoV-2 stimulated gamma interferon release by T-cells [Units/volume] in Blood	36032309
SARS-CoV-2 (COVID-19) IgA Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	586521

Concept Name	Concept Id
SARS-CoV-2 (COVID-19) Ab [Presence] in DBS by Immunoassay	36031197
SARS-CoV-2 (COVID-19) Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	36661369
SARS-CoV-2 (COVID-19) IgM Ab [Presence] in Serum or Plasma by Immunoassay	723475
SARS-CoV-2 (COVID-19) Ab [Interpretation] in Serum or Plasma	723480
SARS-CoV-2 (COVID-19) IgG Ab [Presence] in DBS by Immunoassay	586527
SARS-CoV-2 stimulated gamma interferon release by T-cells [Units/volume] corrected for background in Blood	36031956

Suppl. Table 4

All significant single-condition tests. Listed estimates are odds ratios for the given cohort pre-to-post compared to Controls, and p-values are adjusted across all condition tests for both cohorts (Bonferroni, prior to filtering to significance). Available at https://doi.org/10.5281/zenodo.11188766.

Suppl. Table 5

Summary statistics for patients in the assessment set, with mean and standard deviation of condition era counts in pre- and post-infection phases. Note that the pre-infection phase covers 1 year of patient history, while the post-infection phase covers 6 months post-acute.

Cohort	Life Stage	Phase	Mean # Conditions	SD # Conditions	# Patients	# Sites
Control	adolescent	post	10.296	10.373	10789	32
Control	adolescent	pre	15.76	16.994	10789	32
Control	adult	post	17.518	18.525	180338	34
Control	adult	pre	27.794	29.446	180338	34
Control	pediatric	post	8.894	9.611	16029	32
Control	pediatric	pre	15.815	19.157	16029	32
Control	senior	post	25.357	24.142	95861	33
Control	senior	pre	40.438	36.562	95861	33
COVID	adolescent	post	10.311	12.376	3703	31
COVID	adolescent	pre	15.979	19.2	3703	31
COVID	adult	post	17.177	18.777	60279	34
COVID	adult	pre	28.432	31.169	60279	34

COVID	pediatric	post	10.074	12.872	3724	29
COVID	pediatric	pre	17.001	21.634	3724	29
COVID	senior	post	24.847	24.162	21668	34
COVID	senior	pre	41.522	38.15	21668	34
PASC	adolescent	post	23.287	22.347	66	20
PASC	adolescent	pre	18.893	21.219	66	20
PASC	adult	post	30.281	30.778	2047	32
PASC	adult	pre	34.566	42.242	2047	32
PASC	pediatric	post	21.061	15.282	49	18
PASC	pediatric	pre	19.755	15.492	49	18
PASC	senior	post	42.374	36.527	697	32
PASC	senior	pre	50.292	51.6	697	32

Suppl. Table 6

All topic-level logistic model tests. Estimates are odds ratios for the given cohort and demographic compared to Controls for the same demographic. Ratios where the demographic is listed as NA are for demographic contrasts independent of phase or cohort (model effectiveness checks, see main Methods). P-values are adjusted across all contrast tests (Holm). Available at https://doi.org/10.5281/zenodo.11188766.

Suppl. Table 7

Source ICD code prefixes used to generate Quan-based comorbidity scores.

ICD Prefixes	Charleson Group	Quan Score
'I21','I22','I252'	1: Acute or historical MI	0
'I43','I50','I099','I110','I130','I132','I255','I420','I425','I426','I427','I428','I429','P290'	2: CHF	2
'I70','I71','I731','I738','I739','I771','I790','I792','K551','K558','K559','Z958','Z959'	3: Peripheral vascular disease	0
'G45','G46','I60','I61','I62','I63','I64','I65','I66','I67','I68','I69','H340'	4: Cerebrovascular disease	0
'F00','F01','F02','F03','G30','F051','G311'	5: Dementia	2
'J40','J41','J42','J43','J44','J45','J46','J47','J60','J61','J62','J63','J64','J65','J66','J67','I278','I279','J684','J701','J703'	6: COPD	1
'M32','M33','M34','M06','M05','M315','M351','M353','M360'	7: Rheumatic disease	1
'K25','K26','K27','K28'	8: Peptic ulcer	0
'B18','K73','K74','K700','K701','K702','K703','K709','K717','K713','K714','K715','K760','K762','K763','K764','K768','K769','Z944'	9: Mild liver disease	2
'E100','E101','E106','E108','E109','E110','E111','E116','E118','E119','E120','E121','E126','E128','E129','E130','E131','E136','E138','E139','E140','E141','E146','E148','E149'	10: Diabetes	0

'E102','E103','E104','E105','E107','E112','E113','E114','E115','E117','E122','E123','E124','E125','E127','E132','E133','E134','E135','E137','E142','E143','E144','E145','E147'	11: Diabetes with chronic complications	1
'G81','G82','G041','G114','G801','G802','G830','G831','G832','G833','G834','G839'	12: Paralysis	2
'N18','N19','N052','N053','N054','N055','N056','N057','N250','I120','I131','N032','N033','N034','N035','N036','N037','Z490','Z491','Z492','Z940','Z992'	13: Renal disease	1
'C00','C01','C02','C03','C04','C05','C06','C07','C08','C09','C10','C11','C12','C13','C14','C15','C16','C17','C18','C19','C20','C21','C22','C23','C24','C25','C26','C30', 'C31','C32','C33','C34','C37','C38','C39','C40','C41','C43','C45','C46','C47','C48', 'C49','C50','C51','C52','C53','C54','C55','C56','C57','C58','C60','C61','C62','C63', 'C64','C65','C66','C67','C68','C69','C70','C71','C72','C73','C74','C75','C76','C81', 'C82','C83', 'C84','C85','C88','C90','C91','C92','C93','C94','C95','C96','C97'	14: Localized cancer/leukemia/lymphoma	2
'K704','K711','K721','K729','K765','K766','K767','I850','I859','I864','I982'	15: Moderate/severe liver disease	4
'C77','C78','C79','C80'	16: Metastatic cancer	6
'B20','B21','B22','B24'	17: HIV/AIDS	4