SI material

Supplementary methods

Patient selection and study design

A total of 355 patients were selected from the WP2 Long-term sequelae cohort of the H2020-fundend ORCHESTRA project (Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic)[1, 2, 3, 4, 5]. The different cohorts started independently in early 2020 and the databases and the follow-up were homogenized and standardized. For this study, two ORCHESTRA subcohorts, FrenchCOVID cohort in France and the UNIVR COVID/LC cohort in Italy were included.

The FrenchCOVID cohort (France, clinicaltrials.gov NCT04262921) consisted of hospitalized patients with laboratory confirmed COVID-19. Serum samples were collected between February 2020 and March 2021, daily during hospitalization, at hospital discharge and 3, 6 and 12 months after discharge together with a clinical evaluation. In this cohort, a total of 1327 serum samples from 443 patients were collected. For the current study we selected all patients who provided informed consent, were discharged alive, and had collected at least 1 sample during the follow-up (n=205). We analyzed the following serum samples: earliest available sample during the acute phase, sample upon hospital discharge, samples at 3, 6, or 12 months after the acute phase.

The UNIVR/LC cohort (Italy) included both hospitalized and non-hospitalized patients with COVID-19, who were followed-up for up to 24 months after initial infection. Samples were collected between April 2021 and December 2023, and taken during acute disease, and every 3 months for up to 24 months after the primary infection. Over the entire study period, 7535 serum samples from 2861 patients were collected. Including only patients with at least 1 sample collected during the follow-up, 1235 patients were screened, for a total of 3999 samples. From these, patients were selected for cytokinome analysis with case-control ratio of 2:1. The inclusion criteria for LC patients were the availability of 3 or more samples available, with one collected in the first 6 months (n= 90 were included); or 3 available samples and hospitalization for COVID-19 (n = 11), bringing the total to 101 LC patients. As controls, 49 patients were selected from a total of 744 patients never developed LC, including all patients with at least 4 samples available (n=39). Controls were completed randomly amongst those with at least 3 samples available, one of which collected in the first 6 months of follow-up. The distribution of samples over time is represented in Figure 1.

Clinical definitions

LC was classified according to the WHO definition established by Delphi methodology[6]. In general, LC occurs in individuals with a history of probable or confirmed SARS-CoV-2 infection, usually 3 months from the onset of COVID-19 with symptoms that last for at least 2 months and cannot be explained by an alternative diagnosis. Symptoms may be new onset, following initial recovery from an acute COVID 19 episode, or persist from the initial illness. Symptoms may also fluctuate or relapse over time. The current study adopted the clinical

phenotypes identified by our collaborators before based on principal component analysis, where clinical symptoms were found to cluster together[1]. Briefly, the clinical phenotypes were defined as follows: the respiratory phenotype includes cough and dyspnoea, the chronic pain phenotype includes arthralgia and myalgia, the chronic-fatigue-like phenotype includes headache, fatigue and memory loss and the neurosensorial phenotype includes anosmia and ageusia.

Joint Factor Analysis for partially observed data

Dimensionality reduction of cytokine data was done with Factor Analysis (FA) using the fully observed data matrix. Specifically, we used both FrenchCOVID and UNIVR/LC data before filtering for missing data, allowing for partially observed measurements in the estimation of the latent factors. Let $i = 1, 2, ..., N$ be the number of patients in our joint data, P the number of observed CIBs, and K the number of latent factors, with $K \ll P$. The parametrization of the model is given by:

$$x_i = \Lambda\theta_i + u_i$$

where

x_i is a $P \times 1$ vector of partially observed cytokine markers, Λ is a $P \times K$ factor loadings, θ_i is a $K \times 1$ vector of latent factor scores, and u_i is a vector containing unique errors not captured by the model[7]. Both θ and u are assumed to be normally distributed, with

$$\theta \sim \mathcal{N}(0, I_K)$$
$$u \sim \mathcal{N}(0, \Psi)\,,$$

Where Ψ is a diagonal matrix[7]. From these two distributions, we can derive the following:

$$\Lambda\theta \sim \mathcal{N}(0, \Lambda\Lambda')$$
$$\Lambda\theta + u \sim \mathcal{N}(0, \Lambda\Lambda' + \Psi)\,,$$

Meaning that $p(x)=N(0,\Lambda\Lambda'+\Psi)$ .The joint distribution of the data x and the factor scores θ is then

$$P\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\Lambda\Lambda' + \Psi & \Lambda\\ \Lambda' & I_K\end{bmatrix}\right),$$

With $cov(\theta,x) = \Lambda$. The probability of the data conditional on the factor scores then becomes

$$\begin{aligned} P(x \mid \theta) &= \mathcal{N}(\mathbb{E}[x] + \Lambda I_K(\theta - \mathbb{E}[\theta]), \Lambda\Lambda' + \Psi - \Lambda I_n^{-1}\Lambda') \\ &= \mathcal{N}(\Lambda\theta, \Psi). \end{aligned}$$

Because the joint dataset combines cohorts measuring overlapping but non-identical CIB panels, each subject is observed only on the subset of markers measured in their cohort. Under the assumption that the data are Missing At Random (MAR),the factor loadings Λ and the unique variances Ψ were therefore estimated by Full Information Maximum Likelihood (FIML). FIML maximizes the observed-data log-likelihood by summing case-wise contributions, each evaluated on the subvector of CIBs actually observed for that patient.

Bayesian framework for covariate measurement error correction

Multivariate models were fitted using a Bayesian correction for covariate measurement error following Barlett and Keogh[8] and Schofield[9].

The model is specified hierarchically through three sub-models: 1) an outcome model, 2) a measurement model linking observed cytokines to latent immune factors, and 3) a conditioning model that serves as a prior for the latent factors. First, a logistic regression was used to define the outcome model, which models the relationship between a LC phenotype and latent immune factors:

$$\text{logit}\, P(y_i = 1 \mid \theta_i, Z_i) = \beta_0 + \beta_\theta^T \theta_i + \beta_Z^T Z_i \,,$$

where y_i is a binary phenotype indicator, θ_i is the $K \times 1$ vector of latent immune factor scores for patient i, and Z_i is a $Q \times 1$ vector of clinical covariates measured at acute infection (age, gender, hypertension, diabetes, chronic respiratory disease, and autoinflammatory disease). The parameters β_0, β_θ, and β_Z are the intercept, the log-odds coefficients of the latent factors, and the log-odds coefficients of the clinical covariates, respectively.

Second, the measurement model links observed cytokine values to the latent factors through the factor analysis structure described above:

$$x_i \mid \theta_i \sim \mathcal{N}(\Lambda\theta_i, \Psi) \,,$$

where $\hat{\Lambda}$ and $\hat{\Psi}$ are fixed estimates from the joint FIML factor analysis on the full pooled data, and xi is the $P \times 1$ vector of standardized cytokine measurements for patient i. Third, the conditioning model serves as a prior for the latent factors, conditioning on the clinical covariates but not on the outcome y_i:

$$\theta_i \mid Z_i \sim \mathcal{N}(\alpha_0 + A^T Z_i, \Omega) \,,$$

where α_0 is a $K \times 1$ vector of intercepts, A is a $Q \times K$ matrix of regression coefficients, and $\boldsymbol{\Omega}$ is the $K \times K$ covariance matrix of the latent factors conditional on Z. We decompose $\boldsymbol{\Omega} = \text{diag}(\tau) L L^T \text{diag}(\tau)$, where L is the Cholesky factor of the corresponding correlation matrix and τ is a vector of scale parameters, which improves sampling efficiency.

The joint posterior over all model parameters is given by

$$p(\beta, \alpha_0, A, \Omega, \theta \mid y, x, Z) \propto \prod_{i=1}^{N} p(y_i \mid \theta_i, Z_i, \beta)\, p(x_i \mid \theta_i)\, p(\theta_i \mid Z_i, \alpha_0, A, \Omega)\, p(\beta)\, p(\alpha_0)\, p(A)\, p(\Omega) \,.$$

Critically, θ_i is treated as a random variable rather than a plug-in point estimate. At each MCMC iteration, θ_i is drawn from its full conditional distribution given the cytokine data, the outcome, the covariates, and the current values of all model parameters. This automatically propagates uncertainty in the latent factor scores to the structural coefficients β_θ and β_Z, avoiding the attenuation and confounder-adjustment biases that arise when noisy factor scores are used as fixed regressors[8, 9].

We specified weakly informative priors throughout. For the structural coefficients we used $\beta_0 \sim N(0,5)$, $\beta_\theta \sim N(0,2)$, and $\beta_Z \sim N(0,2)$, corresponding to a 95% prior interval on odds ratios

of approximately (0.02,50). For the conditioning model we used $\alpha_0 \sim N(0,2)$ and $A_{q,k} \sim N(0,2)$. The correlation matrix of Ω was given an LKJ(2) prior on its Cholesky factor, and the scale parameters τ were given half-Cauchy priors with scale 2.5.

Posterior inference was performed using Hamiltonian Monte Carlo via the No-U-Turn Sampler implemented in Stan through the cmdstanr R interface. For each timepoint-outcome combination, we ran 4 chains with 2000 warmup iterations and 2000 sampling iterations each, with adapt_delta set to 0.95 and max_treedepth set to 12 to reduce the risk of divergent transitions and saturated tree depths during sampling. For each parameter, we report a posterior two-sided p-value computed as twice the proportion of posterior draws falling on the opposite side of zero from the posterior mean[10]. Convergence was assessed using three complementary diagnostics computed for every model parameter: the split-$\hat{R}$ statistic (target < 1.01), which compares within-chain and between-chain variance; the tail effective sample size (ESS-tail; target > 400), which quantifies the number of effectively independent draws available; and autocorrelation length (ACL), defined as the ratio of total post-warmup draws to bulk effective sample size[11]. Diagnostic plots for $\hat{R}$, ESS-tail, and ACL across all main-analysis models are reported in SI Figures S6–S8.

Supplementary Tables

Supplementary table S1. Frequency of LC-related symptoms at different timepoints during follow-up. Numbers represent all patients showing symptoms at each respective timepoint upon the total number of patients with clinical data available at the same timepoint. Proportions are indicated in parentheses.

LC symptoms	**3 months *n* = 213/343 (62%)**	**6 months *n* = 210/362 (58%)**	**12 months *n* = 122/250 (48%)**	**18 months *n* = 82/207 (39%)**	**24 months *n* = 21/74 (28%)**
Chronic fatigue-like phenotype	145 (68%)	131 (62%)	78 (64%)	55 (67%)	16 (76%)
Fatigue/malaise	141 (66%)	126 (60%)	71 (58%)	44 (54%)	13 (62%)
Confusion/memory loss	18 (8%)	17 (8%)	12 (10%)	7 (9%)	4 (19%)
Headache	17 (8%)	24 (11%)	18 (15%)	15 (18%)	1 (4.8%)
Respiratory phenotype	90 (42%)	97 (46%)	51 (42%)	31 (38%)	8 (38%)
Cough	35 (16%)	37 (18%)	20 (16%)	11 (13%)	3 (14%)
Dyspnoea/shortness of breath	68 (32%)	74 (35%)	39 (32%)	25 (30%)	7 (33%)
Crackles/chest pain	5 (2%)	0 (0%)	0 (0%)	0 (0%)	1 (4.8%)
Chronic pain phenotype	72 (34%)	82 (39%)	45 (37%)	27 (33%)	6 (29%)
Myalgia	49 (23%)	54 (26%)	29 (24%)	17 (21%)	3 (14%)
Arthralgia	47 (22%)	49 (23%)	34 (28%)	19 (23%)	5 (24%)
Inability to walk	3 (1%)	2 (1%)	3 (2%)	1 (1%)	0 (0%)
Neurosensorial phenotype	32 (15%)	24 (11%)	17 (14%)	15 (18%)	5 (24%)
Anosmia	26 (12%)	19 (9%)	13 (11%)	13 (16%)	5 (24%)
Ageusia	22 (10%)	17 (8%)	12 (10%)	10 (12%)	3 (14%)

Supplementary Figures

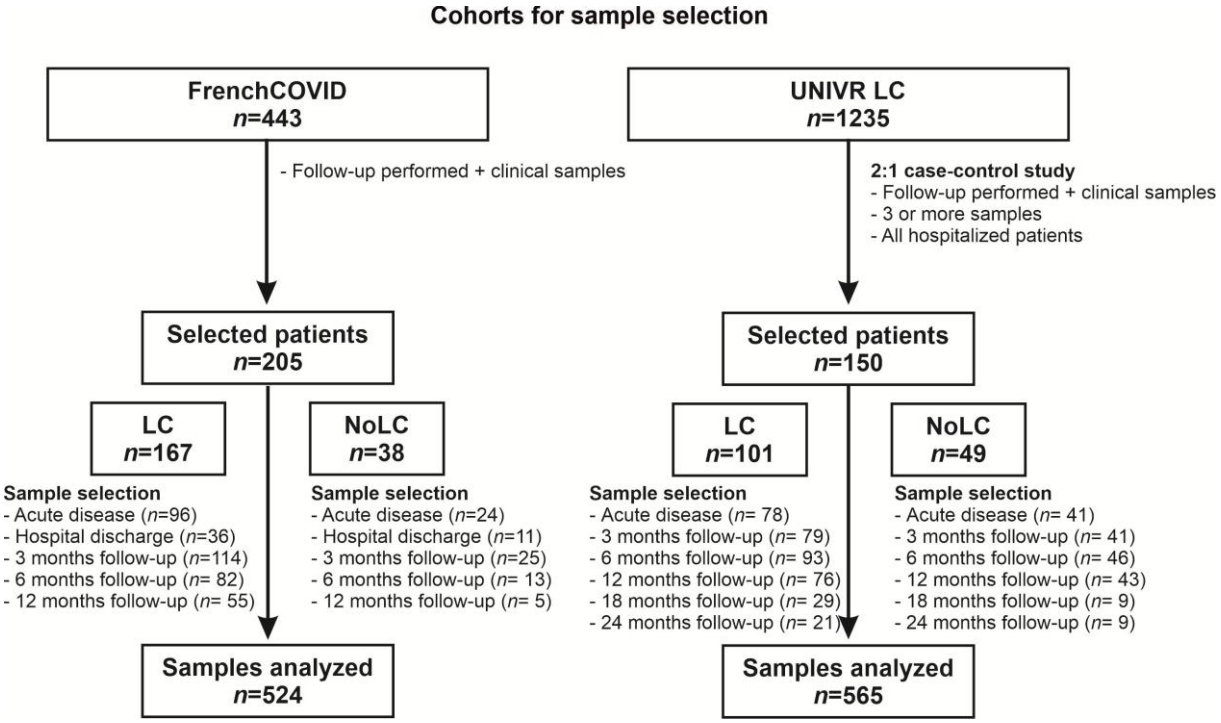

Figure S1. Patients analyzed in the current study were selected from two cohorts in two different countries within the international ORCHESTRA study. The FrenchCOVID cohort was conducted during the early pandemic in France and consists of hospitalized patients infected with pre-omicron viral variants whereas the UNIVR LC subcohort was conducted in Italy during the omicron waves, consisting of ambulatory as well as hospitalized patients. Only patients with the required samples and associated clinical data were considered for this study.

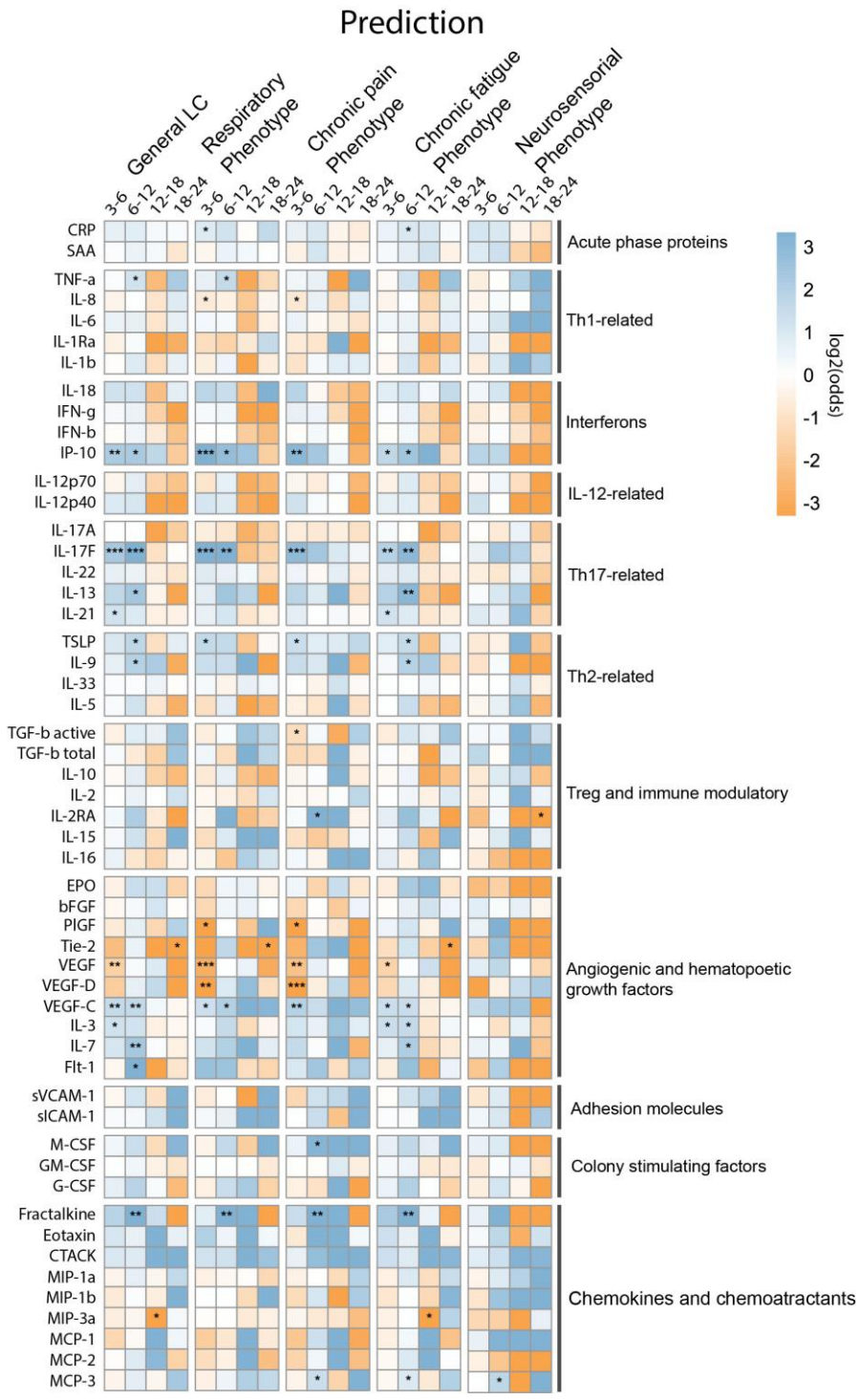

Figure S2. Univariate analysis of circulating immune biomarkers (CIBs) predicting LC phenotypes**.** Colour gradients represent odds ratios (OR) for relative biomarker expression levels, where orange indicates lower concentrations and steel-blue indicates higher concentrations relative to non-LC COVID-19 convalescents. Colour coding is based on log2(odds) with minimum and maximum cut-off at -3 and 3, respectively. Biomarkers are grouped by functional categories: acute-phase proteins, interferons, Th1-, Th2-, and Th17-related cytokines, Treg/immune modulatory factors, angiogenic and hematopoietic growth factors, adhesion molecules, colony-stimulating factors, and chemokines. P-values were calculated using the Wald test. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

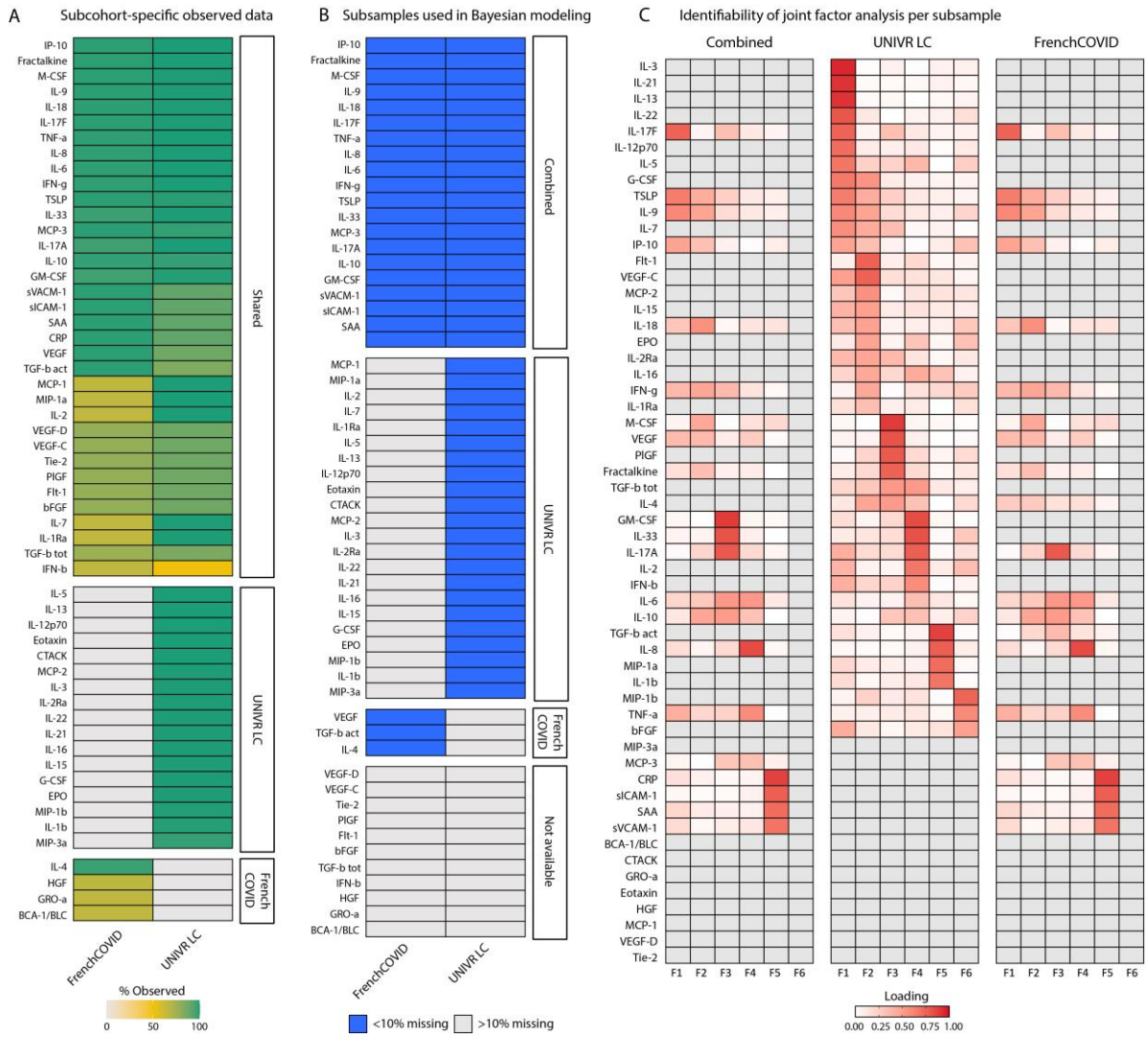

Figure S3. Cytokine data density across cohorts and definition of analytical subsamples. A) Percentage of observed cytokine values for each subcohort and definition of preprocessing marker blocks. B) Data points below the lower limit of detection (LLOD) were imputed using Quantile Regression Imputation for Left Censored Data (QRILC)[12] within the three blocks to avoid introducing artificial constant-value tails in cytokine distributions. C Joint factor loadings by subcohort at the 10% missingness threshold. Heatmap of standardised factor loadings from the joint FIML exploratory factor analysis (6-factor solution), shown separately for each subcohort. Cytokines (rows) are ordered by their primary factor (factor with the highest absolute loading in the full joint solution) and ranked by loading magnitude within each factor. Colour intensity reflects the absolute loading value (white = 0, red = 1). Grey tiles indicate factors that were not retained for a given subcohort because fewer than 3 cytokines exceeded the |0.30| loading threshold; those factors are excluded from the corresponding MESE model. The Combined subcohort includes cytokines passing the 10% missingness filter in both cohorts jointly; the UNIVR LC and FrenchCOVID subcohorts additionally incorporate cytokine-specific extras not available in the combined dataset.

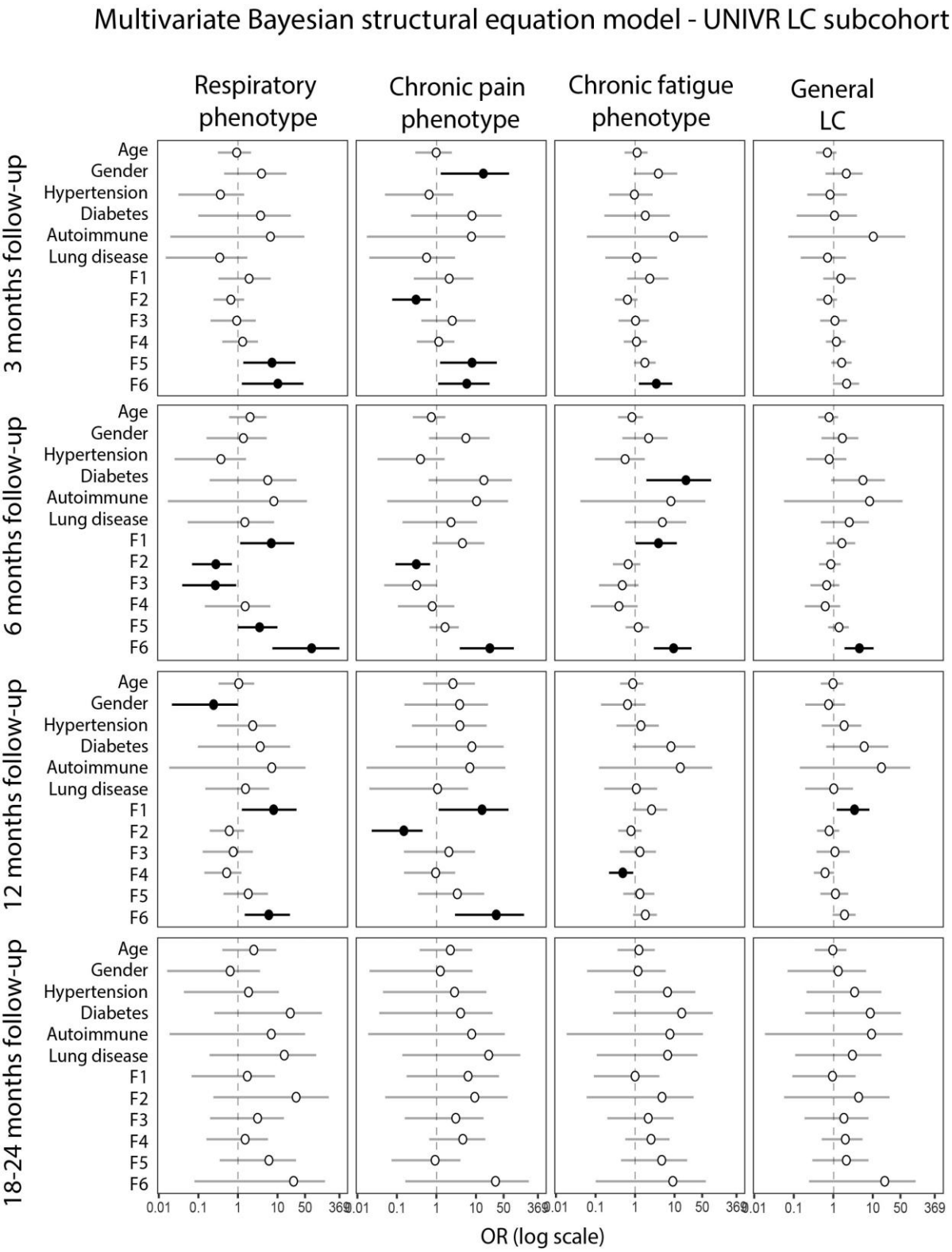

Figure S4. Multivariate Bayesian structural equation model incorporating clinical parameters and factor structures across four longitudinal timepoints (3, 6, 12, and 18–24 months follow-up) of the UNIVR LC subcohort. Data are presented for the three primary LC phenotypes and LC overall.

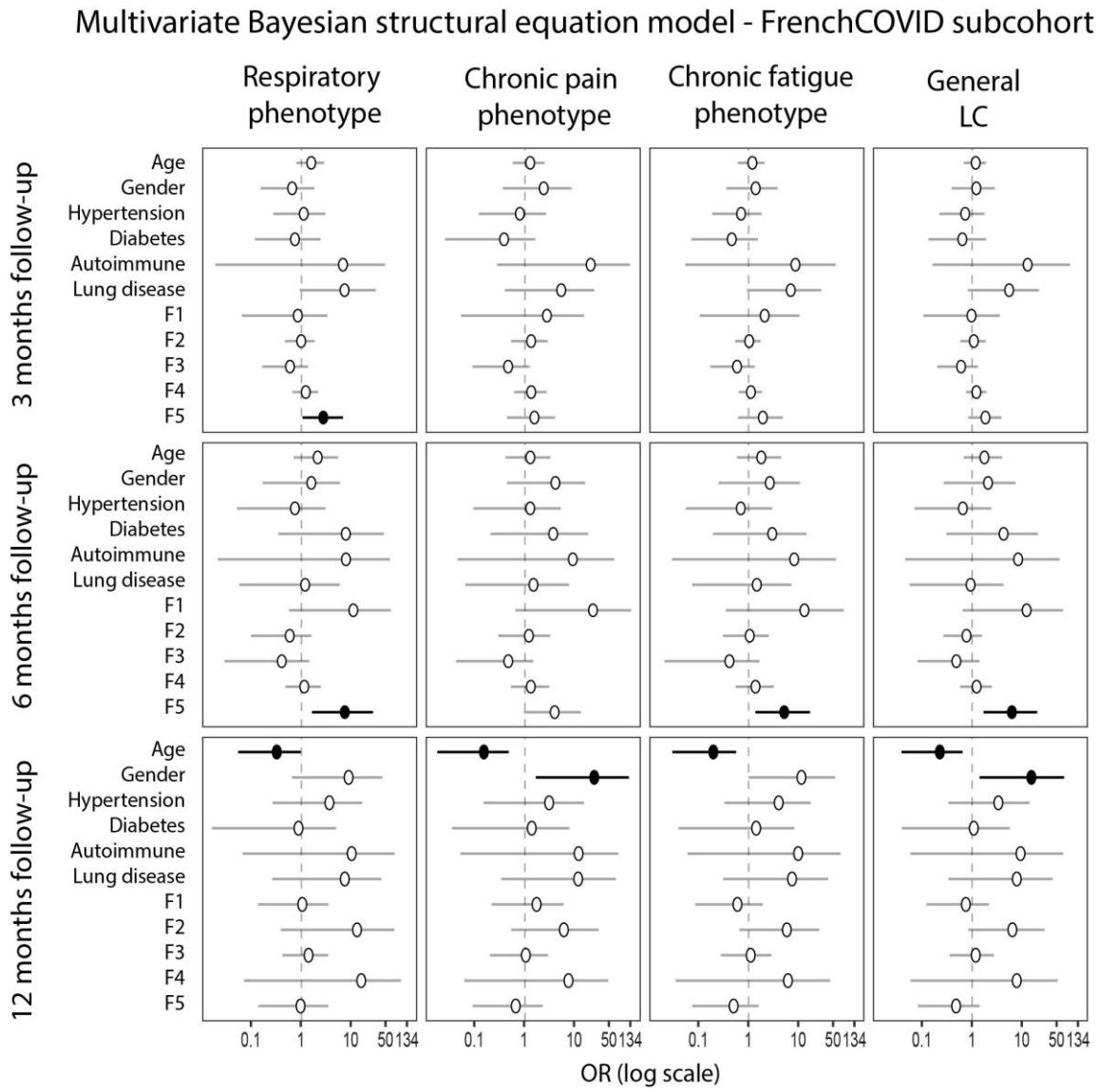

Figure S5. Multivariate Bayesian structural equation model incorporating clinical parameters and factor structures across Three longitudinal timepoints (3, 6, and 12, months follow-up) of the FrenchCOVID subcohort. Data are presented for the three primary LC phenotypes and LC overall.

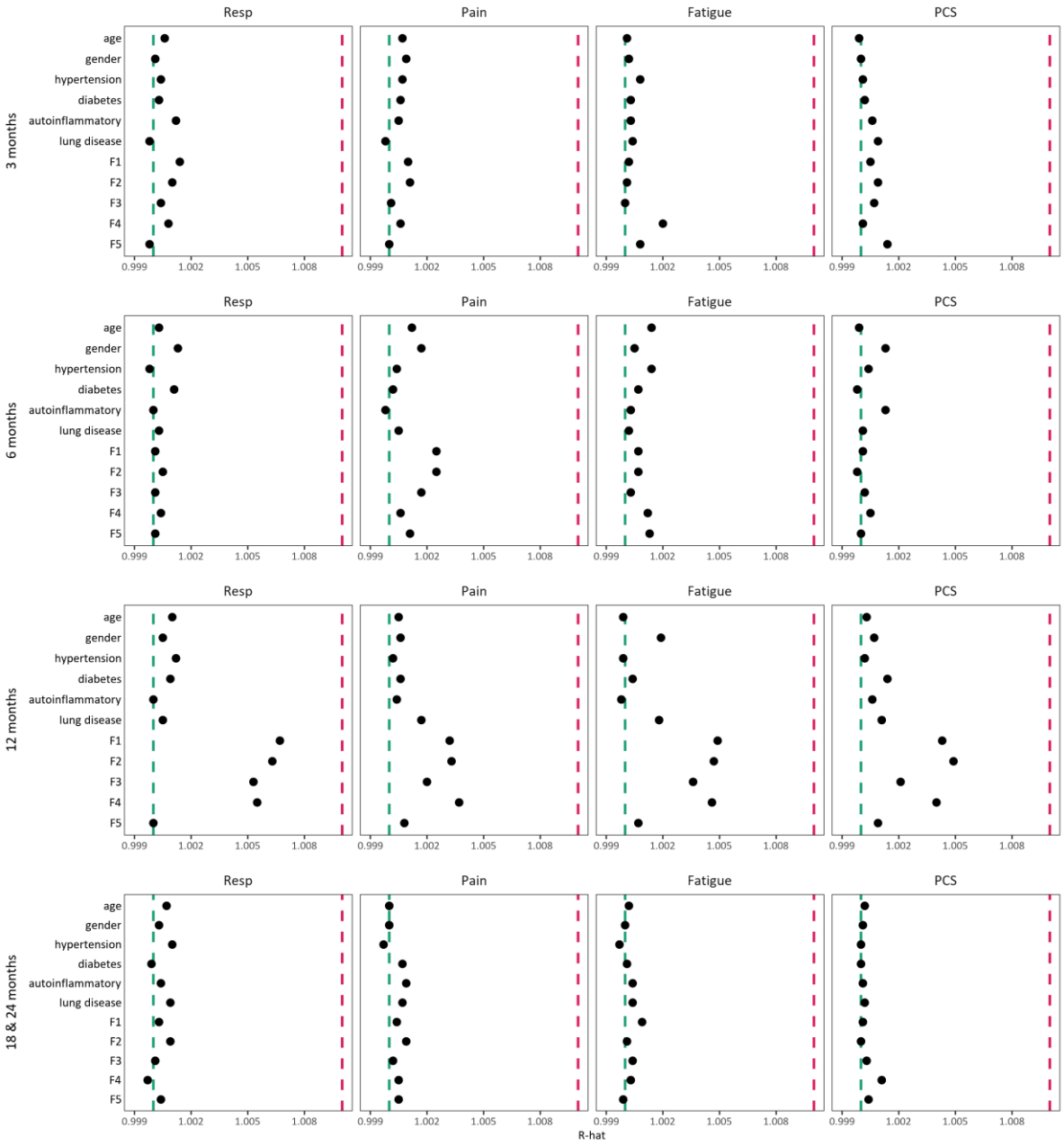

Figure S6. Convergence diagnosis: R-hat. R-hat metrics are shown for all LC phenotypes-timepoints model fits using the combined subset and a missing data threshold of 10%. The green dashed line shows R-hat=1.00 indicating perfect mixing of the four parallel HMC chains, while the red dashed line marks the convergence threshold of R-hat=1.01. All parameters across all models satisfied R-hat<1.01.

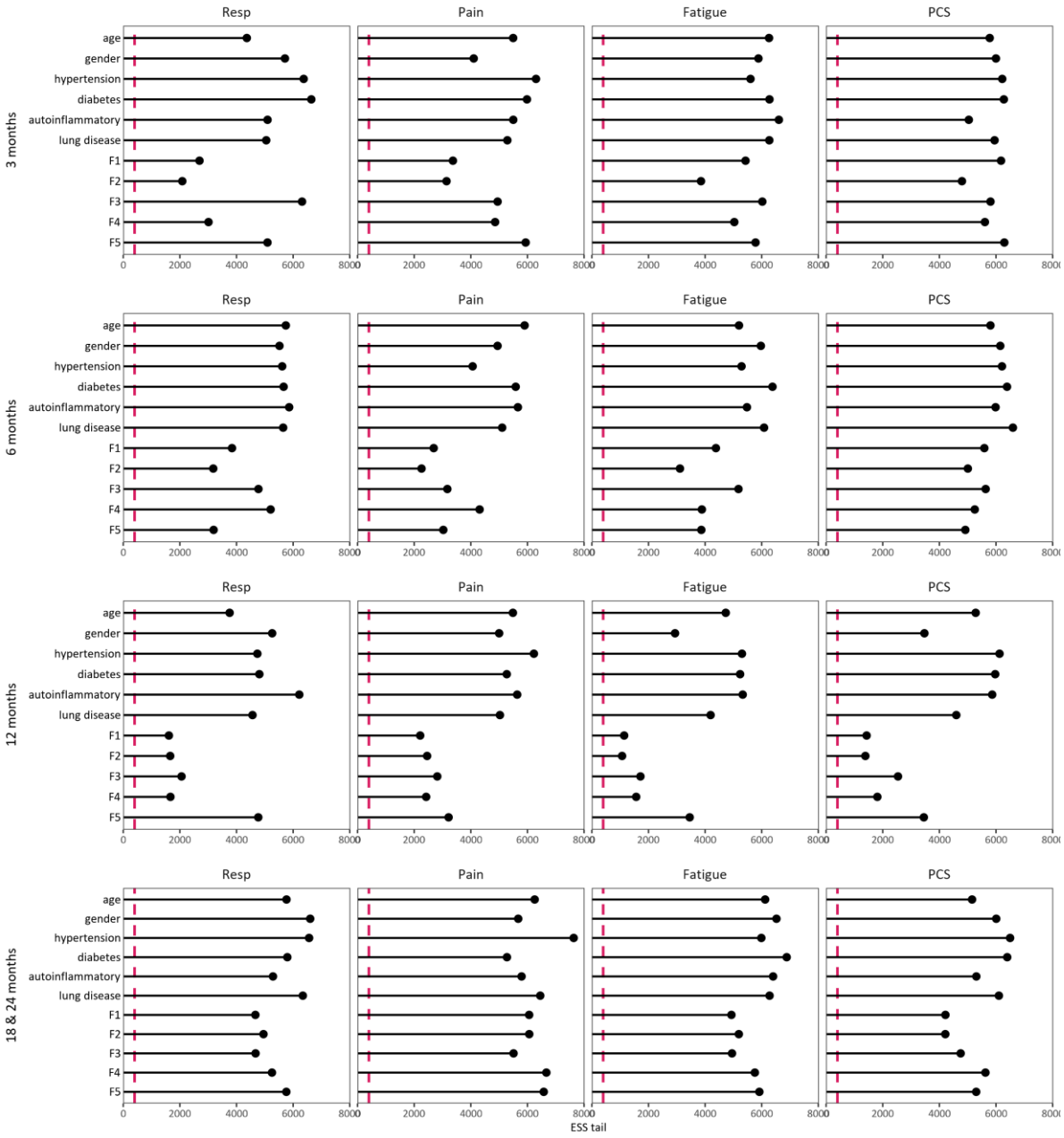

Figure S7. Convergence diagnostics: tail effective sample size (ESS-tail). ESS-tail metrics are shown for all LC phenotypes-timepoints model fits using the combined subset and a missing data threshold of 10%. ESS-tail quantifies the number of effectively independent posterior draws available for estimating the posterior quantiles used to construct 95% credible intervals; higher values indicate more reliable interval estimation. The red dashed line marks the minimum recommended threshold of 400. Across all models, ESS-tail values exceeded this threshold, indicating sufficient effective sample size for posterior inference.

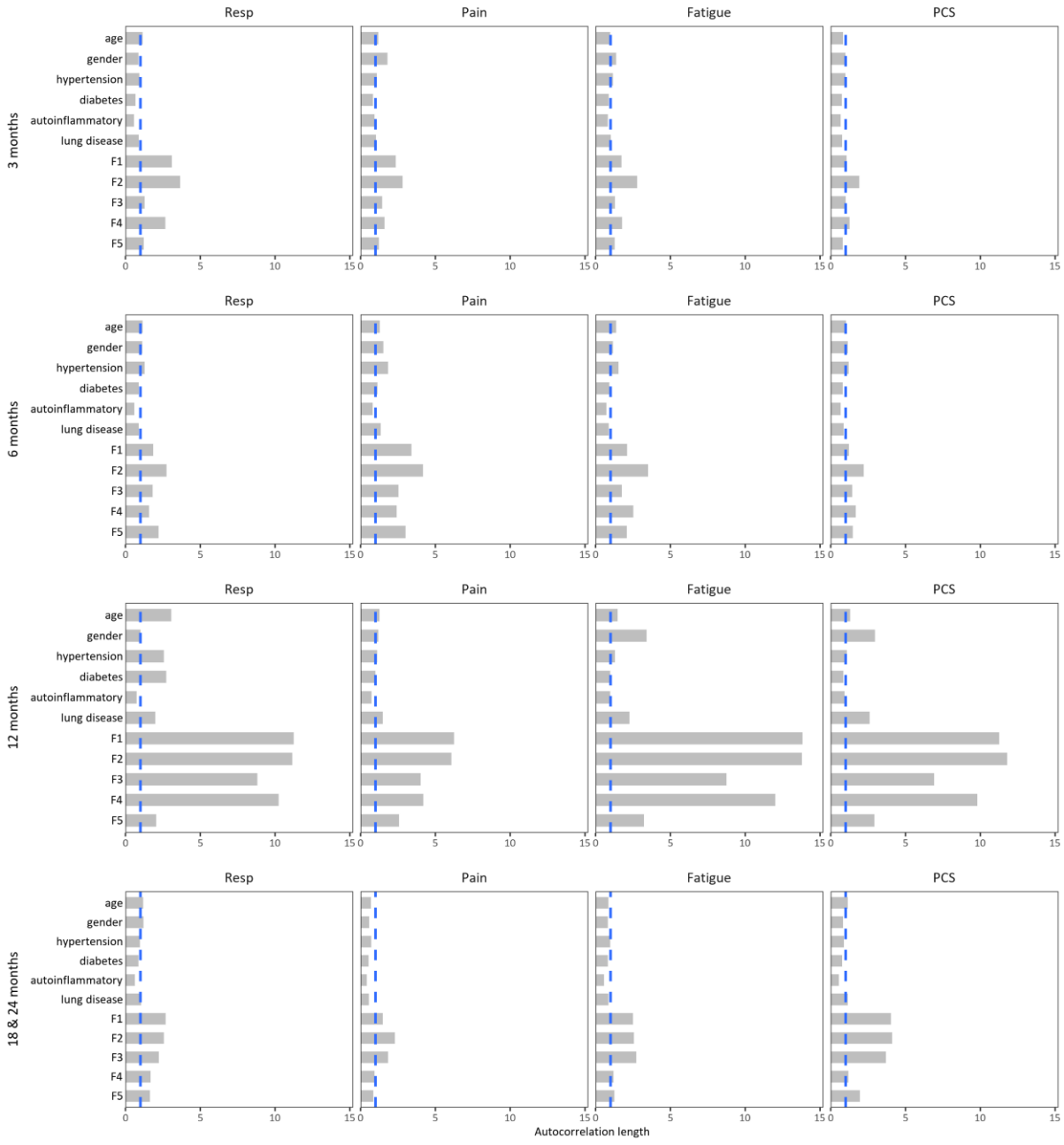

Figure S8. Convergence diagnostics: posterior sample autocorrelation length (ACL). Autocorrelation metrics are shown for all LC phenotypes-timepoints model fits using the combined subset and a missing data threshold of 10%. The blue dashed line marks ACL = 1, the theoretical lower bound corresponding to perfectly independent draws. For models fitted using timepoints 3-6 months, ACL values remained close to 1, indicating low autocorrelation and efficient exploration of the posterior distribution by the HMC sampler. Larger ACL values were observed at 12 months and at the pooled 18/24-month timepoint, reflecting the smaller sample sizes available at later follow-up.

Supplementary references

1. Gentilotti, E. *et al.* Clinical phenotypes and quality of life to define post-COVID-19 syndrome: a cluster analysis of the multinational, prospective ORCHESTRA cohort. *EClinicalMedicine* 62, 102107 (2023).

2. Gentilotti, E. *et al.* Chemosensory assessment and impact on quality of life in neurosensorial cluster of the post COVID 19 syndrome. *Sci Rep* 14, 20951 (2024).

3. Berkell, M. *et al.* Quasi-species prevalence and clinical impact of evolving SARS-CoV-2 lineages in European COVID-19 cohorts, January 2020 to February 2022. *Euro Surveill* 30 (2025).

4. Gusinow, R. *et al.* Latent transition analysis for longitudinal studies of post-acute infection syndromes. *Nat Commun* 17 (2026).

5. Rinaldi, E. *et al.* Harmonization and standardization of data for a pan-European cohort on SARS- CoV-2 pandemic. *NPJ Digit Med* 5, 75 (2022).

6. (WHO), W.H.O. Post COVID-19 condition. [cited May 2026]Available from: https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition

7. Ghahramani F., L.R.J., Willis G.A. Amenability and weak amenability of second conjugate Banach algebras. *Proc. Amer. Math. Soc.* 124, 1489-1497 (1996).

8. Bartlett, J.W. & Keogh, R.H. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. *Stat Methods Med Res* 27, 1695-1708 (2018).

9. Schofield, L.S., Junker, B., Taylor, L.J. & Black, D.A. Predictive Inference Using Latent Variables with Covariates. *Psychometrika* 80, 727-747 (2015).

10. Shi, H., & Yin, G. Reconnecting p-Value and Posterior Probability Under One- and Two-Sided Tests. *The American Statistician* 75, 265–275 (2021).

11. Aki Vehtari, A.G., Daniel Simpson, Bob Carpenter, Paul-Christian Bürkner. Rank-Normalization, Folding, and Localization: An Improvedˆ R for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis, Bayesian Anal.* 2, 667-718 (2021).

12. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci Rep* 8, 663 (2018).