# Methods

## Study Design and Patient Data

The Mass General Brigham (MGB) Biobank contains ~100,000 banked plasma, serum, and DNA samples from >100,000 consented patients. Electronic Medical Record (EMR) data and lifestyle, environment, and family history surveys can also be linked to the banked samples. The Longitudinal EMR and Omics COVID-19 Cohort (LEOCC) consists of a subset of individuals with prospective plasma samples from the MGB Biobank. Patients with a positive COVID-19 diagnosis (defined as a COVID-19 positive infection control flag, COVID-19 presumed infection control flag, or SARS-CoV-2 RNA positive test result) and available plasma samples prior to COVID-19 (up to October 27, 2020) were included. No additional exclusion criteria were applied. Clinical data relevant to COVID-19 infection, including clinical measures, disease diagnoses, and COVID-19 severity were also extracted from EMR data for use in statistical models. This study was approved by the Brigham and Women’s Institutional Review Board (IRB: 2014P001109).

A total of 940 plasma samples from 661 individuals were collected from consented patients and were stored at –80 C. These samples are categorized by the time point of collection relative to a positive COVID-19 diagnosis, including 474 pre-COVID-19 samples (date of collection < date of diagnosis), 282 during COVID-19 samples (collected within 28 days of diagnosis), and 182 post-COVID-19 samples (collected more than 28 days after COVID-19 diagnosis). For patients with multiple during and/or post-COVID-19 samples, only the sample collected at the date closest to diagnosis was retained for during-COVID-19, and only the sample collected at the date furthest from diagnosis was retained for post-COVID-19. Patients without BMI data were also excluded from the sample sets, yielding a total of n = 441 pre-COVID-19, n = 86 during COVID-19, and n = 82 post-COVID-19 samples used for analysis.

The COVID-19 severity level was determined according to WHO guidelines (16) as 0 = ambulatory mild disease (no hospitalization), 1 = hospitalized moderate disease (hospitalized without ventilation), 2 = hospitalized severe disease (hospitalized with ventilation), or 3 = death. Further demographic, risk factor, and comorbidity covariables were defined before COVID-19 diagnosis for all patients as follows: *age* is the numerical patient age; *race* is categorical (black, other, white); *ethnicity* (Hispanic/non-Hispanic), *sex* (female/male), and *smoking* (yes/no) are binary; *BMI* is the numerical median body mass index (BMI) for each patient; and *comorbidity level* is a binary “mild” or “severe” factor based on a Charlson index < 5 or ≥ 5, respectively.

## Charlson Index Calculation

Charlson indexes were calculated as previously described (17) for each patient. Indexes were calculated based on the presence of diseases and age scores listed in Charlson, 2022, where categories were defined by groups of ICD-9 codes and assigned a weight from 1 to 6, where more severe comorbidity categories had a higher weight (18). The sum of the scores of comorbidity presence plus the age score yielded the Charlson index, which is negatively associated with 10-year survival. Charlson indexes were divided into two categories: an index < 5 was defined as *mild* and >= 5 as *severe* comorbidity.

## Metabolomics Analyses

Plasma samples were sent to Metabolon for comprehensive metabolomic profiling of polar and nonpolar metabolite classes in plasma extracts. Samples were extracted and prepared according to methods published previously (19). Quality control (QC) samples were used to control for batch variation throughout the run, using coefficients of variation. Four previously described (20) liquid chromatography-tandem mass spectrometry (LC-MS/MS) methods were used to capture comprehensive metabolite classes, including: 1) amines and polar metabolites (positive ionization), 2) polar metabolites in central metabolism (negative ionization), 3) polar and nonpolar lipids, and 4) lipids and metabolites of intermediate polarity (including free fatty acids and bile acids). Instruments were calibrated each day for mass resolution and mass accuracy, and all materials were purchased from a single lot (21).

Raw metabolite integrals were normalized to the median value (median value = 1) of each metabolite for a given run day. Metabolite identification was performed using automated comparison to features (including retention time, *m/*z, preferred adducts, in-source fragments, and mass spectra) of ~8,000 purified chemical standards in the Metabolon reference library (21). Metabolites that remained unknown, but are reproducibly captured, after comparison with standards were annotated with a unique and persistent “X-number” for potential future identification.

Abundances for a total of 1,546 grouped and aligned metabolites were reported by the Metabolon platform. Metabolites annotated as xenobiotics and with ≥ 75% missing values were removed, yielding 1,108 metabolites in the final dataset. Remaining missing values were imputed as the half-minimum intensity for a given metabolite. Metabolite intensities were log-transformed and *pareto*-scaled for use in regression models.

## Statistical Models

Regression models were applied to assess associations between metabolite levels and (1) COVID-19 severity, where samples were stratified by sample collection time point, and (2) time of sample collection (e.g., pre-, during-, post-COVID-19). Models were adjusted for demographics (age, race, ethnicity, sex), risk factors (BMI, smoking status), and comorbidities (Charlson Index category). Relevant coefficient p-values were adjusted for multiple comparisons using false discovery rates (22) and FDR adjusted p-values < 0.05 were considered statistically significant.

**Equation 1** was used to determine metabolite associations with severity, stratified by pre-, during, and post-COVID-19 samples:

[**Equation 1**]

where *j* is the ordinal COVID-19 severity (levels 0-3) and *metabolite* is normalized and scaled relative metabolite abundance. Modeling was performed using the *polr*() function of the MASS R package (version 7.3-55) (23). From these models, we determined three sets of interest: (1) **predisposition metabolites**, significantly associated with severity in pre-COVID-19 samples only; (2) **predisposition-and-acute metabolites**, significantly associated with severity in the same direction in pre- and during COVID-19 samples; and (3) **persistent metabolites**, significantly associated with severity in the same direction across all times of sample collection.

**Equation 2** was used to determine metabolites associated with time of sample collection (pre, during, and post-COVID-19):

using the *glm*() function in R (version 4.1.3), where *E(Y)* is the normalized and scaled metabolite relative abundance, *time point* is the time of sample collection (pre-, during, and post-COVID-19), and *ε* follows a normal distribution N(0,1).

Results from **Equation 2** were used to identify metabolites that reflect transient alterations during infection and lingering changes after infection. Specifically, we identified the following sets of metabolites based on the statistical significance of the sample collection time term: 1) those altered upon active infection (pre- vs. during COVID-19), 2) metabolites altered after active infection relative to pre-COVID-19 samples (pre- vs. post-COVID-19 samples), and 3) metabolites altered after infection relative to active infection (during vs. post-COVID-19). From these significant metabolites, we defined transient and lingering alterations as follows: **1) transient alterations** during infection include metabolites significantly associated in pre- vs. during COVID-19 and in during vs. post-COVID-19 but not pre- vs post-COVID-19; 2) **lingering alterations** after infection include metabolites significantly associated in pre- vs. post-COVID-19.

## Mapping Metabolites to Pathways

Statistically significant metabolites were mapped to pathways using the getPathwayFromAnalyte() function from RaMP-DB 2.0 (database version 2.0.7, package version 2.3.1) (24) to assist with biological interpretation. Metabolite identifiers (HMDB, PubChem, CAS, and ChemSpider) provided by the Metabolon platform and custom background of all identified metabolites in the entire metabolomics dataset (n = 1,276) were used as input. Only pathways that contained at least two metabolites were retained. Pathways that shared common metabolites/annotations were initially clustered using the findCluster() function. Clusters were then manually inspected and annotated to revise cluster membership.