Full Title: Identifying COVID-19 phenotypes using cluster analysis and assessing their clinical outcomes

Short Title: COVID-19 phenotypes using cluster analysis

Authors: Eric Yamga, MD, MBI¹, Louis Mullie, MD, MSC¹, Madeleine Durand, MD, MSC¹,², Alexandre Cadrin-Chenevert MD³, An Tang MD, MSC²,⁴, Emmanuel Montagnon PhD², Carl Chartrand-Lefebvre MD, MSc²,⁴, Michaël Chassé MD, PhD¹,²

Affiliations

1. Department of Medicine, Centre Hospitalier de l’Université de Montréal, Montreal, Quebec, Canada
2. Centre de recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Québec, Canada.
3. Department of Medical Imaging, CISSS Lanaudière, Université Laval, Joliette, Quebec, Canada
4. Department of Radiology, Radiation Oncology and Nuclear Medicine, Centre Hospitalier de l’Université de Montréal, Montreal, Quebec, Canada

Word count: 3768

Address for Correspondence
Michaël Chassé MD, PhD¹
Department of Medicine
Centre Hospitalier de l’Université de Montréal (CHUM)
1051 Rue Sanguinet, Montréal, QC H2X 3E4
michael.chasse@umontreal.ca

Authors contributions

Eric Yamga : Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing
Louis Mullie : Conceptualization, Data curation, Methodology, Writing – review & editing
Madeleine Durand : Conceptualization, Funding acquisition, Supervision , Writing – review & editing
Alexandre Cadrin-Chenevert : Conceptualization, Data curation, Writing – review & editing
An Tang : Data curation, Funding acquisition, Writing – review & editing
Emmanuel Montagnon : Data curation, Writing – review & editing
Carl Chartrand-Lefebvre : Data curation, Funding acquisition, Writing – review & editing
Michaël Chassé : Conceptualization, Project administration, Funding acquisition, Supervision, Validation, Writing – review & editing
Abstract

Multiple clinical phenotypes have been proposed for COVID-19, but few have stemmed from data-driven methods. We aimed to identify distinct phenotypes in patients admitted with COVID-19 using cluster analysis, and compare their respective characteristics and clinical outcomes.

We analyzed the data from 547 patients hospitalized with COVID-19 in a Canadian academic hospital from January 1, 2020, to January 30, 2021. We compared four clustering algorithms: K-means, PAM (partition around medoids), divisive and agglomerative hierarchical clustering. We used imaging data and 34 clinical variables collected within the first 24 hours of admission to train our algorithm. We then conducted survival analysis to compare clinical outcomes across phenotypes and trained a classification and regression tree (CART) to facilitate phenotype interpretation and phenotype assignment.

We identified three clinical phenotypes, with 61 patients (17%) in Cluster 1, 221 patients (40%) in Cluster 2 and 235 (43%) in Cluster 3. Cluster 2 and Cluster 3 were both characterized by a low-risk respiratory and inflammatory profile, but differed in terms of demographics. Compared with Cluster 3, Cluster 2 comprised older patients with more comorbidities. Cluster 1 represented the group with the most severe clinical presentation, as inferred by the highest rate of hypoxemia and the highest radiological burden. Mortality, mechanical ventilation and ICU admission risk were all significantly different across phenotypes.

We conducted a phenotypic analysis of adult inpatients with COVID-19 and identified three distinct phenotypes associated with different clinical outcomes. Further research is needed to determine how to properly incorporate those phenotypes in the management of patients with COVID-19.
Abbreviations

APN – average proportion of non-overlap
AD – average distance
ADM – average distance between means
CART – classification and regression tree
CCI – Charlson Comorbidity Index
CXR – chest radiographs
FAMD – factor analysis of mixed data
FOM – figure of merit
ICU – intensive care unit
MCI – Medicines Comorbidity Index
MV – mechanical ventilation
NLR – neutrophil-to-lymphocyte ratio
PAM – partition around medoids
PCR – polymerase chain reaction
POLST – Physician Orders for Life-Sustaining Treatment
VIA – variable importance analysis
Introduction

Patients affected with coronavirus disease 2019 (COVID-19) have shown significant clinical heterogeneity and variability in disease trajectory (1). Clinical phenotypes are homogeneous subgroups of a disease presenting distinct clinical features (2). Well-established phenotypes are potentially useful at the bedside for appropriately classifying patients into meaningful categories, predicting disease course, and personalizing treatments.

Since the first case description of COVID-19, various phenotypes have emerged, each using various layers of clinical information. Two phenotypes have been described based on lung mechanics and radiological findings (3,4), others have focused on disease complications—as such, a hypercoagulable phenotype has been observed, prompting recommendations for intensified antithrombotic therapy (5–7).

The majority of those phenotypes failed to fully describe the complexity of the disease as they focused on characterizing one dimension of the clinical presentation. Being mostly derived from clinical observation, or based on outcomes that can only occur in the future is however potentially less relevant in a prospective and real-world context. Thus, the reliability and the methodology of those first phenotyping efforts have been put into question (8,9). Increasing amounts of medical data allow conducting phenotypic analyses using data-driven techniques. Those methods are hypothesis-agnostic and solely rely on the assumption that clinical patterns lie within the data (10). Clustering is an unsupervised machine learning method used to identify homogeneous groups within a heterogeneous dataset. This method has been used to describe clinical phenotypes in other diseases such as asthma (11,12), COPD (13) and sepsis (14).

The primary objective of this study was to identify COVID-19 cluster of phenotypes at patient presentation using multimodal real-world clinical data and medical imaging data (15). Our secondary
objectives were to assess the association of those phenotypes with three clinical outcomes: mechanical ventilation (MV), intensive care unit (ICU) admission and hospital mortality. Finally, we created a simple decision tree classifier allowing the interpretation and assignment of patients to one of the identified phenotypes.

Methods

Data sources

We used real-world data extracted from clinical source system comprising relevant clinical information from all COVID-19-related hospitalizations at the Centre for the Integration and Analysis of Medical Data (CITADEL) of the Centre Hospitalier de l’Université de Montréal (CHUM), a Canadian academic quaternary center. The analytical dataset contains de-identified data for over 1,100 patients hospitalized with COVID-19, including demographics, comorbidities, laboratory results, vital signs, drugs, medical procedures, frontal chest radiographs (CXR) and clinical outcomes. The raw data was managed using SQLite 3, and further data processing was conducted using Python version 3.7 and R version 4.0.3. We provided additional details regarding initial data processing. (see S1 text, supplementary methods).

Study population

We included all unique adult hospitalizations (≥ 18 years of age) for COVID-19 from January 1, 2020, to January 30, 2021, for which a chest X-ray was available within 24 hours of admission. A COVID-19 hospitalization episode was defined as a hospitalization within seven days of a positive SARS-CoV-2 PCR result. The Institutional Review Board of the CHUM (Centre Hospitalier de l’Université de Montréal) approved the study and informed consent was waived because of its low risk and retrospective nature.
Imaging Data Processing

Imaging data were obtained in the DICOM format, and frontal CXR (posteroanterior and anteroposterior) were processed, discarding lateral CXR. Lung opacities observed on CXR were manually annotated with bounding boxes by a board-certified radiologist using a bounding box annotation software (16). This annotation method is recognized by the Radiological Society of North America (RSNA) and is the annotation methodology of choice for all deep learning challenges involving image detection (17,18). Bounding boxes are rectangular or squared delimitations of the opacities found on a given chest radiograph reported with a scaled width and a scaled length. Hence, from the manual annotation, we derived the number of opacities and the total size of opacities as a relative percentage of the total surface of the image.

Variables Selection and Feature Engineering

From the analytical dataset, we extracted 160 candidate variables. We provided the list of those variables in the supplementary material (see S1 Table). For each of those variables, we exclusively used the first recorded value within the first 24 hours of admission. We then excluded 56 variables for which more than 25% of observations were missing. The remaining missing variables were imputed using all available features except clinical outcomes (ICU admission, mechanical ventilation and death). We used classification and regression tree (CART) single mean imputation, a robust method against outliers, multicollinearity and skewed distributions that has the benefit of being simple to implement in a real-world setting (19). We acknowledge that other imputation algorithms such as Expectation Maximization (EM) and Multiple Imputation (MI) have shown superior performance but the additional computational cost makes those approaches difficult to implement in a real-world setting (20).

We then computed three more variables that have recently been associated with COVID-19 mortality: neutrophil-to-lymphocyte ratio (NLR) (21), the ratio of peripheral arterial oxygen saturation to the inspired fraction of oxygen (SpO₂/FiO₂) (22), and the shock index (heart rate/systolic blood pressure)
We also computed the Medicines Comorbidity Index (MCI), a metric to assess multimorbidity that has shown similar epidemiological value to the Charlson Comorbidity Index (CCI) (24). We relied on the MCI instead of the CCI because comorbidities were not systematically recorded in our database, but home medications were (see S1 Table). MCI was computed at the time of study enrollment only using data that was available on presentation to mimic real-world setting. We summarized the final set of variables included for analysis with their respective subdomains of interest in Table 1.

**Table 1. Final set of variables and respective subdomains**

<table>
<thead>
<tr>
<th>Subdomain</th>
<th>Clinical variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>Age, sex, comorbidities (MCI)</td>
</tr>
<tr>
<td>Hemodynamic</td>
<td>SBP, DBP, HR, Shock index (HR/SBP), Anion Gap</td>
</tr>
<tr>
<td>Respiratory</td>
<td>SpO$_2$, SpO$_2$/FiO$_2$</td>
</tr>
<tr>
<td>Imaging</td>
<td>Opacities number, Opacities size</td>
</tr>
<tr>
<td>Hematologic</td>
<td>Neutrophils, Lymphocytes, NLR</td>
</tr>
<tr>
<td>Inflammatory/Thrombotic</td>
<td>MPV, Temperature</td>
</tr>
<tr>
<td>Renal</td>
<td>Creatinine, Sodium, Potassium, Bicarbonate</td>
</tr>
</tbody>
</table>

Despite receiving little attention, feature engineering in clustering efforts has shown to increase learning performance through dimensionality reduction (25). In the healthcare setting, aggregated variables such as the one included in our algorithm also enhances the stability of a learning model’s performance over time (26). In addition to the statistical advantages mentioned, feature engineering also allows for better interpretability, as aggregated variables are often clearer and more intuitive. Pre-processed variables do not impede the hypothesis-agnostic premise of unsupervised machine learning, as all these variables are directly derived from the features of the original dataset and no preferential weighting was conducted.

**Data Preparation and Cluster Analysis**
Before applying clustering algorithms, we adequately processed our dataset, log-transforming skewed continuous variables (skewness > 0.5) and excluding highly correlated variables (correlation > 0.8). The then obtained principal components via factor analysis of mixed data (FAMD) (27) for all our observations. The excluded variables based on high correlation were the total white blood cell count (WBC), the Fraction of Inspired Oxygen (FiO2) and the heart rate (HR). Those three variables were respectively highly correlated with neutrophils, the SpO\textsubscript{2}/FiO\textsubscript{2} ratio and the shock index. We provided additional details regarding the data preparation in the supplementary material (see S1 text).

We compared four clustering algorithms: K-means, PAM (partition around medoids), divisive and agglomerative hierarchical clustering. We used three internal validation metrics (connectivity, Dunn index, the average silhouette width) and four stability measures (average proportion of non-overlap (APN), the average distance (AD), the average distance between means (ADM), and the figure of merit (FOM)) to compare the algorithms. The optimal algorithm and the optimal number of clusters were then determined by rank aggregation of the ranked lists of each validation metric. We used the optCluster package, an R package facilitating the execution of the aforementioned analyses (28).

**Phenotypes Robustness: sensitivity analyses**

We conducted a sensitivity analysis to assess whether the removal of imaging data altered the performance of the clustering algorithm. We first compared the clustering results given these two scenarios using the average silhouette width and the adjusted rand index (29), a measure of the similarity between two data clustering. We then assessed the number of patients that underwent phenotypic reclassification before and after the removal of imaging data in the clustering algorithm. In other words, we analyzed the characteristics of patients for which the cluster assignment differed after the removal of imaging data.

**Statistical Analysis**
Clusters Interpretation

We assessed the difference in the distribution of features across identified clusters using the Chi-Square test for categorical variables, analysis of variance (ANOVA) and Kruskal–Wallis test for normally distributed and non-normally distributed continuous variables respectively.

To facilitate the interpretation and clinical usability of the obtained clusters, we trained a simple decision tree using the CART with the clusters as predicted outcomes (30). We also determined the most critical variables to distinguish the clusters by conducting a variable importance analysis (VIA) (31). Details regarding VIA were provided in the supplementary material (see S1 text, S1 Figure).

Clinical Outcomes Evaluation

We conducted survival analysis using Kaplan-Meier method to compare clinical outcomes according to clusters. We assessed three clinical outcomes: 7-day ICU admission, 7-day mechanical ventilation and 30-day mortality. To reduce confounding bias, we specifically restricted clinical outcomes analysis to patients eligible for MV and ICU admission according to their Physician Orders for Life-Sustaining Treatment (POLST) form.

Results

Study Population

In total, 1,125 unique COVID-19 hospitalizations were screened. A total of 559 patients were excluded after removing readmissions (n = 36), patients without a CXR within 24 hours of admission (n = 523), and patients for which clinical data was missing (n = 19), leaving 547 patients for the cluster analysis (see Fig 1). We provided details regarding the characteristics of our study cohort (see Table 2). Our population was similar to other cohorts of patients hospitalized with COVID-19 in North America during
the first two waves, with a mean age of 69 years old, a relatively similar proportion of men and women, and an in-hospital mortality rate of 20% (32).

Table 2. Baseline characteristics of the study population.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Count (%) or Mean(SD) or Median [IQR]</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>547</td>
</tr>
<tr>
<td>Age (years)</td>
<td>66.56 (17.94)</td>
</tr>
<tr>
<td>Sex (male, %)</td>
<td>313 (57.2)</td>
</tr>
<tr>
<td>Laboratory results</td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>127.23 (20.69)</td>
</tr>
<tr>
<td>Platelet</td>
<td>207.00 [157.00, 271.50]</td>
</tr>
<tr>
<td>WBC</td>
<td>6.80 [5.30,9.85]</td>
</tr>
<tr>
<td>Neutrophil count</td>
<td>5.01 [3.60,7.60]</td>
</tr>
<tr>
<td>Lymphocyte count</td>
<td>0.95 [0.62,1.35]</td>
</tr>
<tr>
<td>Monocytes count</td>
<td>0.63 (0.39)</td>
</tr>
<tr>
<td>Basophils count</td>
<td>0.01 (0.03)</td>
</tr>
<tr>
<td>Eosinophils count</td>
<td>0.05 (0.12)</td>
</tr>
<tr>
<td>VPM</td>
<td>9.89 (1.33)</td>
</tr>
<tr>
<td>VGM</td>
<td>89.90 [85.90,93.85]</td>
</tr>
<tr>
<td>NLR</td>
<td>5.25 [3.19,10.00]</td>
</tr>
<tr>
<td>Sodium</td>
<td>137.59 (5.05)</td>
</tr>
<tr>
<td>Potassium</td>
<td>4.00 (0.53)</td>
</tr>
<tr>
<td>Bicarbonate</td>
<td>24.95 (3.76)</td>
</tr>
<tr>
<td>Anion gap</td>
<td>11.05 (3.69)</td>
</tr>
<tr>
<td>Creatinine (µmol/L)</td>
<td>78.00 [63.00,105.00]</td>
</tr>
<tr>
<td>Vital Signs</td>
<td></td>
</tr>
<tr>
<td>FiO₂</td>
<td>21.00 [21.00,28.00]</td>
</tr>
<tr>
<td>SpO₂</td>
<td>95.00 [94.00,97.00]</td>
</tr>
<tr>
<td>SpO₂/FiO₂</td>
<td>447.62 [339.29,461.90]</td>
</tr>
<tr>
<td>Temperature</td>
<td>36.98 (0.51)</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>130.00 [116.00,144.50]</td>
</tr>
<tr>
<td>Diastolic Blood Pressure</td>
<td>75.00 [68.00,82.00]</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>92.2 (20.2)</td>
</tr>
<tr>
<td>Shock Index</td>
<td>0.69 [0.58,0.82]</td>
</tr>
<tr>
<td>Respiratory Rate</td>
<td>20.00 [20.00,24.00]</td>
</tr>
<tr>
<td>Home Medication</td>
<td></td>
</tr>
<tr>
<td>Medicines Comorbidity Index</td>
<td>2.77 (2.04)</td>
</tr>
<tr>
<td>Anticholesterolemic agents</td>
<td>183 (33.5)</td>
</tr>
</tbody>
</table>
Antihypertensive agents 241 (44.1)
Bronchodilator agents 178 (32.5)
Diuretics 137 (25.0)
Factor Xa Inhibitors 50 (9.1)
Hypoglycemic agents 209 (38.2)
Platelet aggregation inhibitors 150 (27.4)

Imaging Data

<table>
<thead>
<tr>
<th>Opacities Numbers (%)</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>141 (25.8)</td>
</tr>
<tr>
<td>1</td>
<td>89 (16.3)</td>
</tr>
<tr>
<td>2</td>
<td>279 (51.0)</td>
</tr>
<tr>
<td>3</td>
<td>37 (6.8)</td>
</tr>
<tr>
<td>4</td>
<td>1 (0.2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opacities Size (surface area %)</th>
<th>Surface area %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7 [0.0,15]</td>
</tr>
</tbody>
</table>

Clinical outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Count (%) or Mean (SD) or Median [IQR]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of Stay (days)</td>
<td>8.29 [3.48, 18.20]</td>
</tr>
<tr>
<td>Mechanical Ventilation</td>
<td>48 (8.8)</td>
</tr>
<tr>
<td>Wave (1st)</td>
<td>295 (53.9)</td>
</tr>
<tr>
<td>ICU admission</td>
<td>132 (24.1)</td>
</tr>
<tr>
<td>Death</td>
<td>113 (20.7)</td>
</tr>
</tbody>
</table>

Clinical Characteristics of Phenotypes

Agglomerative hierarchical clustering was deemed the most robust clustering algorithm for our dataset. The optimal number of K clusters was K = 3 (see Fig 2), yielding the highest clustering performance as exhibited by the rank aggregation of seven internal validation measures (see S2 Figure, S2 Table).

We summarized the characteristics regarding those clusters in Table 3. As determined through the VIA, the most important variables for discriminating clusters were age, the MCI, the SpO₂/FiO₂ ratio, opacities size, the neutrocyte-lymphocyte ratio, creatinine and the shock index (see Figures 3A-3B).
Table 3. Clinical characteristics stratified by clusters

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (%)</td>
<td>61 (11%)</td>
<td>211 (39%)</td>
<td>275 (50%)</td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>66.18 (15.01)</td>
<td>75.46 (12.54)</td>
<td>59.82 (19.10)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex (male)</td>
<td>26 (42.6)</td>
<td>123 (58.3)</td>
<td>164 (59.6)</td>
<td>0.695</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Laboratory results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>117.69 (21.57)</td>
<td>125.55 (20.83)</td>
<td>130.43 (20.35)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Platelet</td>
<td>244.00 [187.00, 333.00]</td>
<td>195.00 [148.50, 259.00]</td>
<td>207.00 [160.50, 269.00]</td>
<td>0.004</td>
</tr>
<tr>
<td>Neutrophil count</td>
<td>8.70 [5.82, 12.63]</td>
<td>5.03 [3.76, 7.46]</td>
<td>4.63 [3.25, 6.92]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Lymphocyte count</td>
<td>0.60 [0.35, 0.98]</td>
<td>0.90 [0.60, 1.30]</td>
<td>1.06 [0.76, 1.54]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Monocytes count</td>
<td>0.51 (0.31)</td>
<td>0.68 (0.42)</td>
<td>0.61 (0.37)</td>
<td>0.006</td>
</tr>
<tr>
<td>Basophils count</td>
<td>0.01 (0.03)</td>
<td>0.02 (0.03)</td>
<td>0.01 (0.02)</td>
<td>0.358</td>
</tr>
<tr>
<td>Eosinophils count</td>
<td>0.02 (0.04)</td>
<td>0.06 (0.14)</td>
<td>0.04 (0.09)</td>
<td>0.010</td>
</tr>
<tr>
<td>Sodium</td>
<td>137.92 (5.56)</td>
<td>138.08 (5.46)</td>
<td>137.47 (4.58)</td>
<td>0.764</td>
</tr>
<tr>
<td>Potassium</td>
<td>4.24 (0.68)</td>
<td>4.08 (0.54)</td>
<td>3.91 (0.45)</td>
<td>0.013</td>
</tr>
<tr>
<td>Bicarbonate</td>
<td>22.78 (4.85)</td>
<td>25.23 (3.76)</td>
<td>25.20 (3.02)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Anion Gap</td>
<td>13.22 (4.25)</td>
<td>10.67 (3.64)</td>
<td>10.87 (3.43)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Creatinine (µmol/L)</td>
<td>84.00 [62.00, 140.00]</td>
<td>91.00 [71.00, 126.00]</td>
<td>72.00 [58.00, 87.00]</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Vital Signs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FiO₂†</td>
<td>80.00 [40.00, 100.00]</td>
<td>21.00 [21.00, 24.50]</td>
<td>21.00 [21.00, 21.00]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SpO₂</td>
<td>92.00 [88.00, 95.00]</td>
<td>95.00 [94.00, 97.00]</td>
<td>96.00 [94.00, 98.00]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SpO₂/FiO₂</td>
<td>118.75 [92.00, 225.00]</td>
<td>447.62 [368.75, 457.14]</td>
<td>457.14 [442.86, 466.67]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>36.89 (0.36)</td>
<td>36.96 (0.52)</td>
<td>37.00 (0.54)</td>
<td>0.327</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>125.00 [116.00, 153.00]</td>
<td>133.00 [120.00, 149.50]</td>
<td>125.00 [114.00, 140.00]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diastolic Blood Pressure</td>
<td>75.00 (10.67)</td>
<td>73.57 (12.31)</td>
<td>76.34 (13.06)</td>
<td>0.054</td>
</tr>
<tr>
<td>Heart Rate†</td>
<td>91.77 (18.05)</td>
<td>89.61 (17.55)</td>
<td>94.34 (22.33)</td>
<td>0.037</td>
</tr>
<tr>
<td>Shock Index</td>
<td>0.72 [0.54, 0.86]</td>
<td>0.66 [0.58, 0.76]</td>
<td>0.73 [0.61, 0.86]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Respiratory Rate</td>
<td>26.00 [22.00, 32.00]</td>
<td>20.00 [18.00, 24.00]</td>
<td>20.00 [18.50, 23.00]</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Home Medication
**Imaging Data**

<table>
<thead>
<tr>
<th>Opacities Numbers (%)</th>
<th></th>
<th></th>
<th></th>
<th>&lt;0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 (1.6)</td>
<td>34 (16.1)</td>
<td>106 (38.5)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4 (6.6)</td>
<td>47 (22.3)</td>
<td>38 (13.8)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>54 (88.5)</td>
<td>106 (50.2)</td>
<td>119 (43.3)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2 (3.3)</td>
<td>23 (10.9)</td>
<td>12 (4.4)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0 (0.0)</td>
<td>1 (0.5)</td>
<td>0 (0.0)</td>
<td></td>
</tr>
</tbody>
</table>

| Opacities Size (surface area %) | 20 (8) | 9 (7) | 7 (7) | <0.001 |

**Clinical outcomes**

<table>
<thead>
<tr>
<th>Length of stay (days)</th>
<th>14.99 [6.00, 34.42]</th>
<th>9.95 [4.61, 20.04]</th>
<th>5.04 [1.32, 11.51]</th>
<th>&lt;0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical ventilation</td>
<td>24 (39.3)</td>
<td>16 (7.6)</td>
<td>8 (2.9)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ICU admission</td>
<td>43 (70.5)</td>
<td>46 (21.8)</td>
<td>43 (15.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Death</td>
<td>26 (42.6)</td>
<td>51 (24.2)</td>
<td>36 (13.1)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Other**

| Wave (1st) | 38 (62.3) | 122 (57.8) | 135 (49.1) | 0.061 |

Cluster 1 (n = 61, 11%) represented the group of patients with the most severe presentation having the highest NLR (median 13.5), the highest rate of hypoxemia (median SpO₂/FiO₂ 118) and the highest...
radiographic burden, with 91% of patients with at least two pulmonary opacities and a median total opacities size of 20% (ratio of opacities area/total CXR area).

Cluster 2 (n = 211, 39%) and Cluster 3 (n = 275, 50%) were similar regarding the respiratory, hematologic and hemodynamic subdomains. They both presented milder clinical manifestations than Cluster 1, but both phenotypes differed in terms of demographics. Cluster 2 represented the oldest cohort (mean 75.4 years) and Cluster 3 represented the youngest cohort (mean 59.8 years). They also differed by MCI, as Cluster 2 represented the group with the highest proportion of comorbidities, with a mean MCI of 4.30, compared to 1.47 for Cluster 3. Accordingly, Cluster 2 included patients with high proportions of concurrent medication on admission: 68% took anti-hypertensive agents, 63% hypolipemiant agents, 56.9% hypoglycemic agents, 53.6% antiplatelets agents and 37.4% bronchodilator agents.

Phenotypes and clinical outcomes

Among the 547 patients in our study cohort, 436 patients were eligible for ICU admission and MV as deemed by their POLST form, and were thus analyzed for clinical outcomes (see Fig 1).

The cumulative mortality risk was significantly different across clusters (log-rank test, \( p = 0.01 \)). The 30-day mortality risks were respectively 39% (21-53%, 95 CI), 33% (22-43%, 95 CI) and 20% (10-28%, 95 CI) for clusters 1, 2 and 3. The difference in cumulative ICU admission risk and cumulative mechanical ventilation risk were also statistically significant across all clusters (log-rank test respectively, \( p<0.0001 \) and \( p<0.0001 \)). More precisely, the 7-day ICU admission risk was 74% (60-83%, 95 CI) for Cluster 1, 28% for Cluster 2 (20-35%, 95 CI) and 21% (15-26%, 95 CI) for Cluster 3. The 7-day mechanical ventilation risk was 58% (43-70%, 95 CI) for Cluster 1, 14% (8-19%, 95 CI) for Cluster 2 and 11% (6-15%, 95 CI) for Cluster 3 (see Fig 4).

Phenotypes Interpretability and Assignment
We developed a simple decision tree that allows the assignment of patients to their respective clinical phenotypes using the rules shown in Fig 5. Using only four variables (age, MCI, SpO₂/FiO₂, opacities size) and by following between three and five steps, one can assign a patient to one of the three phenotypes with an accuracy varying between 63 and 100% according to the given path.

**Phenotypes Robustness: sensitivity analyses**

When comparing the clustering results with and without the inclusion of imaging data, the adjusted rand index was 0.55 indicating that the similarity of the two clusterings was moderate. The average silhouette width remained unchanged (0.34 with imaging; 0.33 without imaging), indicating that both approaches yielded homogeneous clusters.

We then further characterized the individuals who underwent phenotypic reclassification. 16 percent of patients (n = 87) underwent reclassification (see Fig 6), meaning that their clusters assignment changed after the removal of imaging data. The highest proportion of reclassified observations came from Cluster 1 as 32 percent of observations (n = 20) were reclassified to Cluster 2 or 3 after the removal of imaging data. When analyzing the clinical outcomes of those patients, we noted that 11 out of 20 patients died (55%). Conversely, we noted that eight patients who were initially assigned to Cluster 2 or 3, were reassigned to Cluster 1 (the most severe phenotype) after the removal of imaging data from the algorithm. All of those eight patients survived.

**Discussion**

We identified three clinical phenotypes with distinct clinical characteristics and outcomes using multimodal clinical data in patients admitted with COVID-19. The three phenotypes can be summarized as follows: severely hypoxemic with high radiological burden irrespective of age (Cluster 1), mildly hypoxemic with either a high comorbidity index or old age (Cluster 2), and mildly hypoxemic with a low comorbidity index (Cluster 3).
The identified phenotypes were clinically relevant as they were associated with distinct clinical outcomes. Cluster 1 included patients with the most severe presentation and was thus unsurprisingly the phenotype with the highest mortality and morbidity risk. Cluster 2 and Cluster 3 represented patients with similar milder clinical presentations but with distinct comorbidities profiles. Patients in Cluster 2 had a higher comorbidity index (4.30 vs 1.47) and the 30-day mortality was higher when compared to Cluster 3 (32% vs. 21%). Being able to distinguish phenotypes with appearing similar features, but different outcomes, is clinically meaningful. Those represent patients currently treated identically, but who might benefit from a distinct and targeted treatment approach.

When comparing our results with previous work (33–40), the number of clusters obtained is consistent, as all have identified three phenotypes. However, all cited studies have reported phenotypes suggesting a linear relationship with age and disease severity. This drastically differs from our findings, in which the most severe phenotype (Cluster 1) does not represent the oldest group. In our opinion, our phenotypes reflect the complexity of the distribution of patients with COVID-19 in which age is not the sole determinant of severity. Further research is needed to understand the virological and immunological factors associated with severe infection in this phenotype.

Despite having considered more than 30 variables in our clustering algorithm, only four subdomains were central in establishing the phenotypes: demographics, hematologic features, respiratory features, and imaging data. Socio-demographics, comorbidities (41) and hypoxemia’s (42) impact on the clinical course of patients with COVID-19 have been well documented, and their relative importance in our clustering effort was thus expected. Furthermore, neutrophil-to-lymphocyte ratio (NLR), previously identified as an independent risk predictor for mortality in COVID-19 (43), was accordingly higher in Cluster 1. On the other hand, mean platelet volume (MPV) was not found to significantly impact clustering results despite being associated with severe forms of the disease (44).
Our study emphasizes the importance of imaging data in COVID-19 related clustering efforts. Through our sensitivity analysis, we have shown that the incorporation of CXR enhanced the phenotypes’ clinical value. Even in the absence of ground truth in unsupervised machine learning, we showed that dismissing imaging data reduced the clinical accuracy of our clustering algorithm. Eight patients initially assigned to Cluster 2 or 3 were reclassified to the more severe Cluster 1 after the removal of imaging data from the algorithm. We deemed the initial assignment to the less severe Cluster 2 or 3 as the accurate one, given all eight of those patients survived. Likewise, 20 patients initially assigned to Cluster 1 were reclassified to Cluster 2 or 3 after the removal of imaging data. We deemed the initial assignment to the severe Cluster 1 as appropriate, because the mortality rate of those 20 patients (55%) exceeded those of Cluster 2 and 3 (respectively 24% and 14%). The use of imaging helped properly classify outlier patients: those who had a lower or higher radiological burden than the majority of the patients in their respective phenotypic group (see S3 Figure, S4 Figure).

To our knowledge, opacity size has not previously been used in other COVID-19 phenotyping efforts. Instead, the number of opacities has been used as a proxy variable in only one other paper (33). The number of opacities is generally more accessible as it can directly be extracted from CXR reports, and does not demand manual annotation of medical images. However, studies have shown that even though those two variables provide overlapping information, they are not interchangeable and their respective value differs when it comes to predicting survival and the need for respiratory support in patients with COVID-19 (45). Furthermore, automation of chest X-ray opacities annotation is increasingly feasible with publicly available deep-learning models harmonizing the process (17). We opted for manual annotation in this paper, awaiting further validation of those tools in the COVID-19 population. Nonetheless, chest X-ray annotation should not be viewed as a rate-limiting process. On the contrary, it should be encouraged in healthcare machine learning as it allows uniformizing the inputs during training and upon model deployment (46).
Our study highlights the feasibility and the importance of agnostic approaches to disease phenotyping with no a priori information about patient outcome.

At the bedside, clinical phenotypes help unbiasedly categorize patients. Previous studies have shown that individual risk factors taken alone are insufficient to properly stratify patients with COVID-19 (45). Phenotypes thus offer a simple, and yet holistic means to describing patients with COVID-19 while incorporating both clinical presentation and morbidity risk.

In clinical trials, phenotypes could help harmonizing enrolled participants and facilitate the identification of patients’ subgroups benefiting from a given therapy. Recent clinical trials revealed that distinct clinical presentations mandate distinct treatments, with some therapies showing benefits only in patients with severe disease (47,48). The various inclusion criteria of these trials have made it difficult for clinicians to clearly identify patients who would benefit the most from those novel interventions. The controversy regarding the benefits of anticoagulation in critically and non-critically ill patients is a testimony of this observation (49). Using standardized phenotypes could remove the ambiguous nature of patients subgrouping in future trials and facilitate the comparison of results across trials.

Our study also highlights the clinical usability of clustering-based phenotypes (50). Variables needed to assign patients to phenotypes are readily available at the point-of-care, and cluster assignment can be done following three and five simple steps. Our decision-tree model for cluster assignment would need to be externally validated before being used in the clinical setting, but it nevertheless remains helpful to understand how the three phenotypes differ and how applicable such phenotypes would be in the clinical setting. Moreover, the probabilistic nature of the algorithm allows for direct quantification of the uncertainty, which is an important factor for clinical usability of machine learning algorithms (51). We restricted our clustering effort to the data available within the first 24 hours of admission to ease generalizability and to emulate the timing of triage decision that is often made early upon admission. To
our knowledge, characterizing the evolution of phenotypes over time has not been studied and would mandate further investigation.

We recognize that clinical phenotypes do not offer a comprehensive explanation model for the observed disease heterogeneity (52). However, they lay the groundwork for understanding COVID-19 pathobiology. Studies linking biobank data to clinical phenotypes allow to capture the taxonomic complexity of the disease and describe how phenotypes differ in terms of pathogenic mechanisms (53).

Our study presents some limitations. Multiple variables could not be included because they were either not captured in our electronic health record (e.g., time from onset of symptoms, mechanical ventilation parameters, in-hospital complications) or excluded from our study because of missingness. However, missing values are common in clinical practice and investigating risk stratification while considering the inherent characteristics real-world data is of importance at the bedside (54). In addition, this enhances the applicability of our phenotypes, as they are only based on the most common variables available for patients admitted with COVID-19 (55). This differs from studies that have included flux cytometry and CD4+/CD8+ count in their algorithm (34). Besides, those omitted variables do not seem to have had significant impact on our results as the three clusters obtained were consistent in numbers with previous work (33–39).

Additionally, our study included patients admitted between January 1, 2020, and January 31, 2021, being before the approval of the majority of targeted therapies against COVID-19 or vaccination. We therefore did not assess the effect of vaccination, treatments and the type of variant on phenotypes. Accordingly, this put our algorithm at risk for temporal dataset shift (56) and calibrating our clustering algorithm will be necessary before exploiting it in the clinical setting. Finally, because race-based data is not recorded in the Quebec healthcare system (57), we could not proceed to sensitivity analysis according to race. For that
very same reason, we acknowledge that our work could be subject to algorithmic bias as evidence has shown racial disparities in clinical outcomes of patients with COVID-19 (58).

**Conclusion**

We developed a multidimensional phenotypic analysis of patients with COVID-19 and identified three distinct phenotypes, with one specifically associated with worse clinical outcomes. Our study supports the feasibility of using real-world clinical data to conduct unsupervised phenotypic clustering. External validation and further research are needed to determine how phenotypes could impact clinical trials design and phenotypic-guided treatment in clinical practice.
Acknowledgments

M.C. is supported by a Fonds de Recherche Québec Santé (FRQS) Clinical Research scholarship. M.D. is supported by a FRQS Clinical Research scholarship. A.T. is supported by a FRQS Clinical Research Scholarship and a Fondation de l'Association des Radiologistes du Québec (FARQ) Clinical Research Scholarship. This work was possible with the contribution of the CHUM Center for Integration and Analysis of Medical Data (CITADEL).

This work was partly funded by a Quebec Bio-Imaging Network research grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reproducibility

The entire code excluding the dataset is publicly available on GitHub (https://github.com/CODA-19/models/tree/master/phenotyper).

The data that support the findings of this study are available on request from the corresponding author, MC.

Disclosures

None

Supporting information

S1 text (Supplementary Methods)

S1-S3 Tables

S1-S4 Figures
References

20. Duy Le T, Beuran R, Tan Y. Comparison of the most influential missing data imputation


58. Bodenreider O, Peters LB, Nguyen T. RxClass-Navigating between Drug Classes and RxNorm
Drugs. ICBO. 2014;106.


63. Arora S, Hu W, Kothari PK. An Analysis of the t-SNE Algorithm for Data Visualization. 2018 Jul 3;


Tables

Table 1. Final set of variables and respective subdomains

Table 2. Baseline characteristics of the included population

Table 3. Clinical characteristics stratified by clusters

Figures

Figure 1. Study Inclusion and Exclusion Criteria

Figure 2. Clusters visualization on factor analysis of mixed data principal components plot

Figure 3A. Radar Plot showing the distribution of clinical variables across clusters

Figure 3B. Bar Plot showing the distribution of clinical variables across clusters

Figure 4. Kaplan-Meier curves of clinical outcomes stratified by clinical outcomes (death, ICU admission and mechanical ventilation)

Figure 5. Clusters assignment decision tree-based rules

Figure 6. Sankey diagram assessing clustering stability with and without imaging data
Figures

Fig 1. Study Inclusion and Exclusion Criteria
Fig 2. Clusters visualization on factor analysis of mixed data principal components plot
Fig 3A. Radar Plot showing the distribution of clinical variables across clusters
**Fig 3B.** Bar Plot showing the distribution of clinical variables across clusters

In both figures, the variables were scaled using normalization. Non-continuous variables underwent log-transformation if the skewness was > 0.5 and the respective mean of each scaled variable was plotted.
Fig 4. Kaplan-Meier curves for clinical outcomes (a) death, (b) ICU admission, and (c) mechanical ventilation stratified by phenotypes.

A.
B.

Log-Rank

\[ p < 0.0001 \]

Cumulative ICU admission

Days

Strata: cluster=1, cluster=2, cluster=3
C.

```
Log-Rank
p < 0.0001

Cumulative Mechanical Ventilation

0.00 0.25 0.50 0.75 1.00
0 5 10 15 20 25

Days

Strata  cluster=1  cluster=2  cluster=3
```
Fig 5. Clusters assignment through decision tree-based rules

This graph represents the classification rules obtained after training a CART decision tree algorithm with the phenotypes as outcomes. The rules obtained could be used at the bedside to determine the phenotype of a given individual following simple steps.

The rules are located in the white squared boxes and each colored node box displays the probability of the predicted class.
Fig 6. Sankey diagram assessing clustering stability with and without imaging data

This plot shows the distribution of the patients that underwent reclassification after the removal of imaging data in the clustering algorithm. 87 observations (16%) were reclassified when ignoring imaging data. Although the clusters are relatively stable, a disproportionate number of reclassified observations originate from Cluster 1, as 20 observations (32%) were reclassified after the removal of imaging. As Cluster 1 is the most severe phenotype, the potential impact of this reclassification is not without consequences.
Initial patient population

All COVID-19 episodes between January 1, 2020 and January 30, 2021 (n=1125)

Unique COVID-19 hospitalizations (n=1089)

Excluded (n=36)
- 36 readmissions removed

Excluded (n=523)
Patients without CXR order within 24 hours of admission

Excluded (n=19)
Patients with CXR within 24 hours of admission missing from the dataset

Final study population for cluster analysis (n=547)

Final study population for outcomes analysis (n=436)
Patients eligible to MV and ICU admission based on POLST form