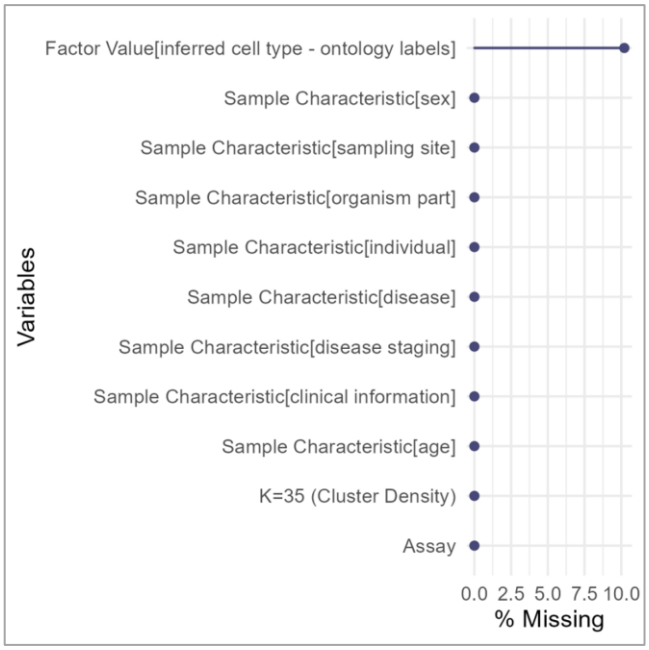
Supplementary A, B, C, D

Supplementary A- Missing Data Analysis

1. Model used in this Study

Logistic regression was used to determine the nature of the missing data (MAR, MNAR or MCAR)- MCAR: Missing completely at random, MAR: Missing at random, or. MNAR: Missing not at random. This logistic regression model was used to investigate the association between the likelihood of missing data ('missing indicator') and various sample characteristics (i.e. input columns). Significant predictors of missingness include the disease status (normal vs. COPD), with normal status being less likely to have missing data, and two levels of smoking intensity, indicating that certain clinical information is related to data availability. Notably, some coefficients and standard errors are not defined due to singularities, suggesting collinearity or insufficient variation within those predictors (only 1 value in column, for example). Overall, the model reveals that both disease status and specific smoking habits significantly influence the pattern of missing data in this dataset. So, there is no evidence of data being MCAR, as significant predictors of missingness have been identified (two used later in H1, H2, H3). The analysis supports the conclusion that the data is MAR, given that missingness is associated with measured variables. There is no direct evidence from this summary that the missingness depends on unobserved data (MNAR), although this cannot be completely ruled out without further analysis specifically designed to detect MNAR.



2. Applying Logistic Regression

Logistic regression can be used to determine the nature of missing data by analysing patterns in the data and examining the relationship between missingness and other variables. Here's how it can be applied to identify whether the data is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR):

2.1. Identifying MCAR (Missing Completely at Random)

- **MCAR Hypothesis:** The missingness is completely unrelated to any observed or unobserved data.
- **Approach:** Perform a logistic regression where the dependent variable is a binary indicator of whether a value is missing (1 if missing, 0 if observed). The independent variables are all the observed variables.

- **Interpretation:** If none of the independent variables are significantly associated with the missingness indicator, the data can be considered MCAR.

2.2. Identifying MAR (Missing at Random)

- **MAR Hypothesis:** The missingness is related to the observed data but not the unobserved data.
- **Approach:** Again, perform a logistic regression with the missingness indicator as the dependent variable and the observed variables as independent variables.
- **Interpretation:** If one or more observed variables are significantly associated with the missingness indicator, the data may be MAR.

2.3. Identifying MNAR (Missing Not at Random)

- **MNAR Hypothesis:** The missingness is related to the unobserved data.
- **Approach:** This is more complex to test because it involves unobserved data. You can start by examining the residuals from the MAR model. If there are patterns in the residuals that suggest systematic missingness, this could indicate MNAR.
- **Techniques:**
    - Sensitivity Analysis: Vary assumptions about the missing data mechanism and examine how the results change.
    - Pattern Analysis: Look for patterns in the data that might suggest a relationship between missingness and unobserved data.
    - Compare different imputation methods and their results.

3. Practical Steps for Logistic Regression Analysis

1. Create Missingness Indicator Variables:
    - For each variable with missing data, create a binary indicator variable that denotes whether the data is missing (1) or observed (0).
2. Perform Logistic Regression:
    - Use the binary missingness indicators as the dependent variables in separate logistic regression models.
    - Use the available observed variables as predictors in these models.
3. Analyse Results:
    - Assess the significance of the predictors.
    - If predictors are significant, the data may be MAR.
    - If no predictors are significant, the data may be MCAR.
    - Investigate further if MNAR is suspected, possibly using more advanced techniques like pattern mixture models or selection models.

Supplementary B – Consolidated Table of Cell Count Per Cell Type Calculations.

Cell Count Per Cell Type Analyses												
Cell Type	Total cells per cell type Non-Smoker N=1	Percentge of total cells per cell type (non-smokers)	Total cells per cell type - Smokers N=3	Percentage of Total cells per cell type (smokers)	Average cells per cell type - Smokers N=3 (devided by 3 samples)	Total Cells per Cell type	Difference in Cell count per cell type - Non-Smokers V Smokers	Total cells per cell type - Smokers without COPD N=2	Average cells count per cell type - Smokers without COPD N=2	Cell count per cell type (COPD) Smoker N=1	Cell count per cell type (Normal) Smokers N=2; Non-smoker N=1	Average Cell count per cell type (Normal) Smokers N=2; Non-smoker N=1
B cell	4	0.11%	268	7.52%	89.3	272	-85.3	267	134	1	271	90
fibroblast of lung	46	1.29%	29	0.81%	9.7	75	36.3	23	12	6	69	23
lung ciliated cell	51	1.43%	124	3.48%	41.3	175	9.7	88	44	36	139	46
lung endothelial cell	37	1.04%	23	0.64%	7.7	60	29.3	18	9	5	55	18
lung macrophage	310	8.69%	870	24.40%	290.0	1180	20.0	612	306	258	922	307
lung secretory cell	11	0.31%	117	3.28%	39.0	128	-26.0	94	47	23	105	35
lymphocyte	3	0.08%	63	1.77%	21.0	66	-18.0	50	25	13	53	18
mast cell	42	1.18%	195	5.47%	65.0	237	-23.0	180	90	15	222	74
natural killer cell	5	0.14%	51	1.43%	17.0	56	-12.0	28	14	23	33	11
T cell	13	0.36%	165	4.63%	55.0	178	-42.0	102	51	63	115	38
transformed epithelial cell	165	4.63%	15	0.42%	5.0	180	160.0	5	3	10	170	57
type I pneumocyte	13	0.36%	88	2.47%	29.3	101	-16.3	52	26	36	65	22
type II pneumocyte	1	0.03%	493	13.83%	164.3	494	-163.3	214	107	279	215	72
(blank)	52	1.46%	312	8.75%	104.0	364	-52.0	208	104	104	260	87
Total	753		2813			3566				872	2694	
											3566	
Hypothesis 1: Compare total cells per cell type - COPD v Non-COPD (Normal) - Compare Column R (pink) with T (Orange)			Hypothesis 2: Compare total cells per cell type - Smokers without COPD (normal) with Non-Smokers Column I (Blue) with Q (Yellow).			Hypothesis 3: Compare total cells per cell type - All Smokers v Non-Smokers Compare Column I (Blue) with M (Green)						

Table 2: Consolidated Table of Cell count Per Cell Type Calculations; Mean; Cell Count variance per category: Percentage of total cells per cell type per category.

Supplementary C – Cell Type Classification

1. **B Cell**:
    - **Function**: A type of white blood cell that plays a crucial role in the adaptive immune system by producing antibodies. They can differentiate into plasma cells that secrete antibodies and memory B cells that provide long-term immunity.
    - **Location**: Found in the bone marrow, lymph nodes, spleen, and blood.
2. **Fibroblast of Lung**:
    - **Function**: A type of cell that synthesizes the extracellular matrix and collagen, providing structural support to tissues. In the lung, fibroblasts are involved in tissue repair and fibrosis.
    - **Location**: Distributed throughout the connective tissue of the lung.
3. **Lung Ciliated Cell**:
    - **Function**: A type of epithelial cell that has hair-like structures called cilia. These cilia beat in a coordinated manner to move mucus and trapped particles out of the respiratory tract.
    - **Location**: Lining the airways of the respiratory tract, particularly in the trachea and bronchi.
4. **Lung Endothelial Cell**:
    - **Function**: A type of cell that forms the inner lining of blood vessels in the lung. They are involved in the exchange of gases, nutrients, and waste products between the blood and lung tissue.
    - **Location**: Lining the blood vessels within the lungs.
5. **Lung Macrophage**:
    - **Function**: A type of immune cell that engulfs and digests cellular debris, foreign substances, and pathogens. They play a critical role in immune defense and inflammation.
    - **Location**: Found in the alveoli (alveolar macrophages) and interstitial tissue of the lungs.
6. **Lung Secretory Cell**:
    - **Function**: A type of epithelial cell that secretes mucus and other substances to protect and lubricate the respiratory tract.
    - **Location**: Found in the airway epithelium, including goblet cells and Clara cells.
7. **Lymphocyte**:
    - **Function**: A type of white blood cell involved in the adaptive immune response. Lymphocytes include B cells, T cells, and natural killer cells.
    - **Location**: Found in the blood, lymphatic system, and lymphoid organs such as the spleen and lymph nodes.
8. **Mast Cell**:
    - **Function**: A type of immune cell that contains granules rich in histamine and heparin. They play a key role in allergic reactions and inflammation.
    - **Location**: Found in connective tissues throughout the body, including the lungs.

9. **Natural Killer Cell**:
    - **Function**: A type of lymphocyte that can kill tumour cells and virus-infected cells without the need for prior sensitization. They play a role in innate immunity.
    - **Location**: Found in the blood and various lymphoid tissues.
10. **T Cell**:
    - **Function**: A type of lymphocyte that plays a central role in cell-mediated immunity. They can be further classified into helper T cells (CD4+), cytotoxic T cells (CD8+), and regulatory T cells.
    - **Location**: Found in the thymus (where they mature), blood, and lymphoid tissues.
11. **Transformed Epithelial Cell**:
    - **Function**: Epithelial cells that have undergone genetic changes, often leading to uncontrolled growth and cancer.
    - **Location**: Can occur in various epithelial tissues, including those in the lung.
12. **Type I Pneumocyte**:
    - **Function**: A type of alveolar cell that forms the thin barrier (alveolar-capillary membrane) for gas exchange in the lungs.
    - **Location**: Lining the alveoli of the lungs.
13. **Type II Pneumocyte**:
    - **Function**: A type of alveolar cell that produces and secretes surfactant, a substance that reduces surface tension in the alveoli and prevents lung collapse.
    - **Location**: Scattered among the Type I pneumocytes in the alveoli.

Supplementary D – Chi-Square Test for Statistical Significance

The Chi-squared (Chi^2) value, also known as the chi-square statistic, is a measure of the discrepancy between the observed and expected frequencies in a contingency table. Here's what it signifies:

1. **Chi^2 Value:**
    - It quantifies the difference between the observed counts (from your data) and the counts that would be expected if there were no association between the variables (under the null hypothesis).
    - The formula for calculating the Chi value is:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency and E is the expected frequency for each cell in the table.

    - A higher Chi^2 value indicates a larger discrepancy between observed and expected frequencies, suggesting that there may be a significant association between the variables.

2. **Interpreting the Chi2 Value:**
    - **Low Chi^2 Value:** Suggests that the observed data fits the expected data well. In this case, you are more likely to accept the null hypothesis, indicating no significant association between the variables.
    - **High Chi^2 Value:** Suggests that the observed data does not fit the expected data well. In this case, you are more likely to reject the null hypothesis, indicating a significant association between the variables.

3. **p-Value:**
    - The p-value associated with the Chi^2 value helps determine the statistical significance of the result.
    - A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, leading to its rejection.
    - A large p-value (> 0.05) indicates weak evidence against the null hypothesis, leading to its acceptance.

Summary of Results *From the Chi^2 tests performed on data:*

1. **Hypothesis 1: COPD vs Non-COPD**
    - **Chi^2 Value:** 320.144
    - **p-Value:** 2.74×10−612.74×10−61
    - Interpretation: There is a significant association between the cell type distribution in parenchymal lung tissue of individuals with COPD and those without COPD.
2. **Hypothesis 2: Smokers vs Non-Smokers**
    - **Chi2 Value:** 499.611

- **p-Value:** 2.68×10−992.68×10−99
- Interpretation: There is a significant association between the cell type distribution in parenchymal lung tissue of individuals without COPD who smoke and those who do not smoke.

3. **Hypothesis 3: Smokers with vs without COPD**
   - **Chi2 Value:** 693.869
   - **p-Value:** 9.05×10−1419.05×10−141
   - Interpretation: There is a significant association between the cell type distribution in parenchymal lung tissue of individuals with or without COPD who smoke, and those who do not smoke.

In all cases, the high Chi2 values and extremely small p-values indicate that the null hypotheses (no significant association) are rejected, confirming significant differences in the distribution of cell types between the compared groups.

Interpretation

For all three hypotheses, the p-values are extremely small (much less than the typical significance level of 0.05). This indicates that there are significant differences in the distribution of cell types between the compared groups in each hypothesis:

1. There is a significant association between the cell type distribution in parenchymal lung tissue of individuals with COPD and those without COPD.
2. There is a significant association between the cell type distribution in parenchymal lung tissue of individuals without COPD who smoke and those who do not smoke.
3. There is a significant association between the cell type distribution in parenchymal lung tissue of individuals who smoke (with or without COPD) and those who do not smoke.

In all cases, the null hypotheses are rejected, meaning there are significant differences in cell type distributions between the compared groups associated with Smoking and Disease Status.

Meaning in Context:

1. **Extremely Small Probability:**
   - 2.2×10^{-16} represents 0.00000000000000022.
   - Such a small p-value indicates that the observed data is extremely unlikely under the null hypothesis.
2. **Statistical Significance:**
   - A p-value this small is much lower than the conventional significance levels (e.g., 0.05, 0.01, or even 0.001).
   - It provides very strong evidence against the null hypothesis.
3. **Context in Chi-Squared Test:**
   - When you see a p-value of 2.2×10−16 in the output of a chi-squared test, it implies that the differences between the observed and expected frequencies are highly significant.

- In practical terms, it means that the null hypothesis (no association) can be rejected with extremely high confidence.

Usage Note:

- The 2.2×10^{-16} value is often the smallest p-value reported by many statistical software packages due to precision limits. It indicates that the actual p-value is smaller than this threshold.

The critical value for a chi-square test depends on the chosen significance level (alpha) and the degrees of freedom (df). For a degree of freedom (df) of 12, the critical values for commonly used significance levels are as follows:

- **For alpha = 0.05 (95% confidence level):** The critical value is approximately 21.026.
- **For alpha = 0.01 (99% confidence level):** The critical value is approximately 26.217.
- **For alpha = of 0.001 (99.9% confidence level)**: The critical value is approximately 32.909.