

Exploratory Data Analysis Report

Dataset Overview

- Rows: 1000
- Columns: 10
- Columns: nctId, briefTitle, officialTitle, conditions, overallStatus, studyType, sex, minimumAge, maximumAge, eligibilityCriteria
- All columns are text/categorical in the raw data.

Data Quality

Missing Values

- officialTitle: 10 missing
- minimumAge: 69 missing
- maximumAge: 488 missing
- Other columns: 0 missing

Duplicates

- Duplicate rows: 0

Column Summaries (Categorical)

overallStatus (top categories)

- COMPLETED: 579
- UNKNOWN: 144
- RECRUITING: 102
- TERMINATED: 59
- NOT_YET_RECRUITING: 41
- WITHDRAWN: 35
- ACTIVE_NOT_RECRUITING: 31
- ENROLLING_BY_INVITATION: 7
- SUSPENDED: 2

studyType

- INTERVENTIONAL: 778
- OBSERVATIONAL: 222

sex

- ALL: 867
- FEMALE: 87
- MALE: 46

conditions (top categories)

- Healthy: 22
- Prostate Cancer: 8
- Breast Cancer: 8
- Multiple Sclerosis: 7
- Stroke: 7
- HIV Infections: 7
- Obesity: 5
- Lung Cancer: 4
- COVID-19: 4
- Chronic Obstructive Pulmonary Disease: 4

minimumAge (top values)

- 18 Years: 620
- 20 Years: 30
- 40 Years: 23
- 50 Years: 20
- 12 Years: 18
- 65 Years: 15
- 16 Years: 14
- 21 Years: 13
- 19 Years: 12
- Missing: 69

maximumAge (top values)

- 65 Years: 60
- 75 Years: 56
- 80 Years: 53
- 70 Years: 30
- 55 Years: 26
- 45 Years: 24
- 85 Years: 24
- 40 Years: 20
- 50 Years: 18
- Missing: 488

Notes

Interpretation of Missingness

- `maximumAge` is missing in nearly half the records, which suggests many trials do not report an upper age limit.
- `minimumAge` is mostly present, indicating lower age bounds are more consistently reported.
- `officialTitle` has minimal missingness (1%), which is unlikely to affect analysis.

Interpretation of Key Distributions

- `overallStatus` is dominated by COMPLETED trials, which implies the dataset is skewed toward finished studies rather than ongoing recruitment.
- `studyType` is heavily INTERVENTIONAL, so results may not reflect observational research equally.
- `sex` is mostly ALL, meaning most trials include both sexes; sex-specific trials are relatively rare in this sample.
- `conditions` are diverse; no single condition dominates, indicating broad clinical coverage.

Limitations

- No numeric columns exist in raw form, so numeric distributions and correlations are not directly applicable.
- Age fields are strings (for example, "18 Years"), preventing numeric age summaries without preprocessing.
- `eligibilityCriteria`, `briefTitle`, and `officialTitle` are free-text fields; summary statistics here require text-based analysis rather than numeric EDA.

Additional Steps Performed

Parsed Age Fields (numeric years)

- `minimumAge` and `maximumAge` were parsed into numeric years by extracting the first numeric token.
- Parsed summary (counts reflect non-missing numeric values):
- min_age_years: count 931, mean 20.14, min 0, 25% 18, median 18, 75% 18, max 75
- max_age_years: count 512, mean 57.14, min 2, 25% 40, median 65, 75% 75, max 120

Outlier Checks (IQR method)

- `minimumAge`: 311 outliers flagged because $Q1 = Q3 = 18$ (distribution heavily concentrated at 18).
- `maximumAge`: 0 outliers flagged by IQR.

Basic Data Cleaning Summary

- Trimmed whitespace and standardized case for `overallStatus`, `studyType`, and `sex` to uppercase.
- Unique counts before vs. after cleaning were unchanged for all columns, indicating minimal whitespace/case issues.

Recommended Next Steps

- Parse `minimumAge` and `maximumAge` into numeric years to enable histograms, boxplots, and summary stats.
- Normalize and split `conditions` if multiple conditions are stored in one cell to improve frequency analysis.
- Perform text analytics on `eligibilityCriteria` and titles (length stats, keyword frequencies, or topic modeling).