STUDY DATASET: PE-EHR[1]

This is de-identified data for the PE-EHR+ study, which is a study to validate efficient tools to identify patients with pulmonary embolism from electronic health records.

Explanation: This analysis of the PE-EHR+ dataset seeks to utilise electronic health record (EHR) data to clarify the complexities of pulmonary embolism (PE) diagnosis and care. By exploring data patterns, we can identify age and demographic variations in PE incidence, as well as trends in comorbidities that affect thromboembolic risk. These findings will support better risk stratification, inform targeted prevention strategies, and deepen our understanding of systemic factors contributing to disparities in diagnosis and treatment. Ultimately, this work aims to leverage evidence to promote equitable approaches that improve outcomes for diverse patient populations affected by PE.

Research Questions:

1. Are there significant differences in PE incidence rates across different age groups (e.g., <40, 40-60, 60-75, >75 years)?
2. How do comorbidity burdens differ between PE-positive and PE-negative patients?
3. Are there racial or ethnic disparities in treatment approaches during hospitalisation?

Objective 1: The Dataset in the virtual PyCharm environment is set up with my GitHub repository for the project:
https://github.com/TanmayeeKodali/PE-EHR-Study

Objective 2: Summary statistics for all teh variables in the dataset

COMPREHENSIVE SUMMARY STATISTICS: ALL VARIABLES

Variable	Overall N={len(df)}	PE Negative N={len(pe_neg)}	PE Positive N={len(pe_pos)}
DEMOGRAPHICS			
Age Groups (Original Categories), n (%)			
18-25 years	51 (3.0%)	31 (3.7%)	20 (2.3%)
25-35 years	159 (9.3%)	110 (13.0%)	49 (5.7%)
35-45 years	145 (8.5%)	76 (9.0%)	69 (8.0%)
45-55 years	213 (12.4%)	101 (11.9%)	112 (13.0%)
55-65 years	369 (21.5%)	165 (19.4%)	204 (23.6%)
65-75 years	400 (23.4%)	189 (22.3%)	211 (24.4%)
75-85 years	255 (14.9%)	118 (13.9%)	137 (15.9%)
85-95 years	113 (6.6%)	53 (6.2%)	60 (7.0%)
95+ years	7 (0.4%)	6 (0.7%)	1 (0.1%)
Age Categories (Collapsed), n (%)			
<40	355 (20.7%)	217 (25.6%)	138 (16.0%)
40-60	582 (34.0%)	266 (31.3%)	316 (36.6%)
60-75	400 (23.4%)	189 (22.3%)	211 (24.4%)
>75	375 (21.9%)	177 (20.8%)	198 (22.9%)

Sex, n (%)			
Female	896 (52.3%)	473 (55.7%)	423 (49.0%)
Male	816 (47.6%)	376 (44.3%)	440 (51.0%)
Race/Ethnicity, n (%)			
Non-Hispanic White	1368.0 (79.9%)	675.0 (79.5%)	693.0 (80.3%)
Non-Hispanic Black	160.0 (9.3%)	71.0 (8.4%)	89.0 (10.3%)
Non-Hispanic Asian	50.0 (2.9%)	30.0 (3.5%)	20.0 (2.3%)
Hispanic/Latinx	288.0 (16.8%)	138.0 (16.3%)	150.0 (17.4%)
American Indian or Alaska Native	4.0 (0.2%)	3.0 (0.4%)	1.0 (0.1%)
Native Hawaiian or Other Pacific Islander	1.0 (0.1%)	0.0 (0.0%)	1.0 (0.1%)
Non-Hispanic Other or Unknown	84.0 (4.9%)	51.0 (6.0%)	33.0 (3.8%)
Other races/ethnicities	67.0 (3.9%)	41.0 (4.8%)	26.0 (3.0%)
Unknown/Declined race/ethnicity	39.0 (2.3%)	22.0 (2.6%)	17.0 (2.0%)
COMORBIDITIES			
Comorbidities Present, n (%)			
Diabetes	57.0 (3.3%)	29.0 (3.4%)	28.0 (3.2%)
Hypertension	940.0 (54.9%)	453.0 (53.4%)	487.0 (56.4%)
History of CVA	121.0 (7.1%)	59.0 (6.9%)	62.0 (7.2%)
Coronary Artery Disease	316.0 (18.5%)	151.0 (17.8%)	165.0 (19.1%)
Heart Failure	291.0 (17.0%)	158.0 (18.6%)	133.0 (15.4%)

Dialysis	26.0 (1.5%)	14.0 (1.6%)	12.0 (1.4%)
Prior VTE History	276.0 (16.1%)	158.0 (18.6%)	118.0 (13.7%)
COVID-19 (Prior 30 days)	52.0 (3.0%)	17.0 (2.0%)	35.0 (4.1%)
Peripheral Artery Disease	148.0 (8.6%)	71.0 (8.4%)	77.0 (8.9%)

PE DIAGNOSIS CHARACTERISTICS

Discharge Diagnosis Group, n (%)

Group-1	568 (33.2%)	45 (5.3%)	523 (60.6%)
Group-2	568 (33.2%)	230 (27.1%)	338 (39.2%)
Group-3	576 (33.6%)	574 (67.6%)	2 (0.2%)

PE Subtypes (Among PE+ Patients Only), n (%)

Subsegmental PE only	N/A	N/A	104.0 (12.2%)
Cor pulmonale present	N/A	N/A	359.0 (41.6%)

DIAGNOSTIC PROCEDURES PERFORMED

Procedures Performed, n (%)

CT Pulmonary Angiography (CTPA)	1468 (85.7%)	114 (13.4%)	620 (71.8%)
Chest CT	1340 (78.2%)	252 (29.7%)	418 (48.4%)
Ventilation-Perfusion Scan	86 (5.0%)	27 (3.2%)	16 (1.9%)
Pulmonary Angiography	2 (0.1%)	0 (0.0%)	1 (0.1%)
Radiology Consultation	372 (21.7%)	61 (7.2%)	125 (14.5%)

Additional Procedures Count (excluding CTPA)

Mean (SD)	0.83 (17.25)	0.33 (0.49)	0.50 (0.53)
Median [IQR]	0.0 [0.0-1.0]	0.0 [0.0-1.0]	0.0 [0.0-1.0]

Excessive Diagnostics (>1 additional procedure), n (%)

Yes	21 (1.2%)	9 (1.1%)	11 (1.3%)

COR PULMONALE DETECTION METHODS (Among 359 PE+ patients with cor pulmonale)

Detected on CT	247.0 (70.2%)
Detected on Echocardiography	153.0 (46.5%)
Troponin elevation	243.0 (70.8%)

Objective 3: Statistical Analysis Plan (SAP)

Aim 1: Age Group and PE Incidence

As we explore the data for this aim, we will first examine the distribution of patients across the four age categories to ensure adequate sample sizes (minimum n=30 per group). We will create frequency tables using pandas.crosstab() to visualize the distribution of PE diagnosis across age groups and identify any sparse cells that might violate chi-square test assumptions. We will also calculate preliminary PE incidence rates for each age group to understand the magnitude of differences before formal testing.

Statistical Approach:

To address the first aim that seeks to determine whether PE incidence rates differ across age groups, a Pearson's chi-square test of independence will be employed using scipy.stats.chi2_contingency(). This test will compare PE diagnosis rates (yes/no) across four age categories (<40, 40-60, 60-75, >75 years). If any expected cell frequencies are less than 5, Fisher's exact test (scipy.stats.fisher_exact()) will be used instead.

Following a significant omnibus test, pairwise comparisons between age groups will be conducted using 2×2 chi-square tests with Bonferroni correction (adjusted $\alpha = 0.05/6 = 0.0083$) to control for multiple comparisons.

Model to Create and Test:

A multivariable logistic regression model will be fit using statsmodels.api.Logit() with PE diagnosis as the outcome and age category as the predictor (using age <40 as the reference group). Odds ratios with 95% confidence intervals will be generated from this analysis to quantify the strength of association between each age group and PE diagnosis.

Tables and Figures:

- Table 1: Crosstabulation of age groups × PE status with row/column percentages
- Table 2: PE incidence rates by age group with odds ratios and 95% CIs
- Figure 1: Grouped bar chart showing PE incidence rates by age group, stratified by hypertension status (addresses both Aim 1 and Aim 2)

Aim 2: Hypertension and PE Association (Stratified by Age)

We will begin by examining the prevalence of hypertension overall and within each PE status group using pandas.DataFrame.groupby() and .value_counts(). We will create stratified contingency tables for each age group to assess whether there are sufficient cases in each cell (hypertension × PE status within each age stratum) for valid statistical testing. We will look for potential effect modification by visually inspecting whether the hypertension-PE relationship appears to strengthen or weaken across age groups. This preliminary exploration will guide our decision on whether to include an interaction term in the regression model.

Statistical Approach:

For the second aim that seeks to assess whether hypertension is associated with PE diagnosis, and whether this relationship varies by age, a two-part analytical approach will be employed.

Univariate Analyses:

First, an overall chi-square test (using scipy.stats.chi2_contingency()) or Fisher's exact test (scipy.stats.fisher_exact()) if cell counts <5 will evaluate the association between hypertension status (yes/no) and PE diagnosis (yes/no) in the full sample. An odds ratio with 95% confidence interval will be calculated using statsmodels.stats.contingency_tables.Table2x2().

Second, to test whether the hypertension-PE association differs across age groups, age-stratified chi-square tests will be conducted. We will loop through each of the four age groups (<40, 40-60, 60-75, >75 years) using Python, constructing separate 2×2 contingency tables and calculating age-specific odds ratios with 95% confidence intervals. This stratified analysis will reveal whether hypertension's effect on PE diagnosis is consistent across the lifespan or varies by age.

Model to Create and Test:

A multivariable logistic regression model will be fit using statsmodels.api.Logit() to formally test for effect modification. The model will include PE diagnosis as the outcome, with hypertension, age category, and sex as predictors, plus an interaction term (Hypertension × Age Category). A significant interaction ($p < 0.10$) would provide statistical evidence that age modifies the hypertension-PE relationship. The model equation is:

$$\text{logit}(PE) = \beta_0 + \beta_1(\text{Hypertension}) + \beta_2(\text{Age\_40-60}) + \beta_3(\text{Age\_60-75}) + \beta_4(\text{Age\_>75}) + \beta_5(\text{Sex}) + \beta_6(\text{HTN}\times\text{Age\_40-60}) + \beta_7(\text{HTN}\times\text{Age\_60-75}) + \beta_8(\text{HTN}\times\text{Age\_>75})$$

Tables and Figures:

- Table 3: Overall hypertension prevalence by PE status with odds ratio
- Table 4: Age-stratified associations showing hypertension-PE odds ratios for each age group
- Figure 1 (same as Aim 1): Grouped bar chart showing PE incidence by age and hypertension status - visually displays the interaction

Aim 3: Racial/Ethnic Disparities in Diagnostic Intensity (PE+ Patients Only)

We will begin by filtering the dataset to include only PE-positive patients (df[df['PE_in_index_hospitalization'] == 1]) and examining the sample size for each racial/ethnic group using pandas.DataFrame['Race_Ethnicity'].value_counts(). Groups with very small sample sizes (n<10) may need to be collapsed into an "Other/Unknown" category. We will create a new variable called "Additional_Procedures" by summing CT_chest, VP_scan, and PulmonaryAngio (excluding CTPA, which is the gold standard). We will examine the distribution of this count variable using histograms (matplotlib.pyplot.hist()) and summary statistics to determine if it is normally distributed or skewed. We will also assess whether the data show overdispersion (variance > mean), which would indicate the need for negative binomial regression rather than Poisson. Additionally, we will create a binary "Excessive_Diagnostics" variable (1 if Additional_Procedures > 1, else 0) and calculate the overall prevalence of excessive diagnostics.

Statistical Approach:

For the third aim that seeks to identify racial/ethnic disparities in diagnostic procedure intensity, analyses will be restricted to patients with confirmed PE diagnosis (PE_in_index_hospitalization = 1).

Diagnostic intensity will be operationalized in two ways:

- Additional Procedures Count: Sum of non-CTPA diagnostic procedures (CT chest + V/Q scan + pulmonary angiography), range 0-3
- Excessive Diagnostics: Binary indicator (1 = patient received more than one additional diagnostic procedure beyond CTPA, 0 = otherwise)

Univariate Analyses:

To compare the mean number of additional procedures across racial/ethnic groups, we will first test for normality within each group using scipy.stats.shapiro()

and test for homogeneity of variances using scipy.stats.levene(). Based on these results, we will either use one-way ANOVA (scipy.stats.f_oneway()) if assumptions are met, or the Kruskal-Wallis H test (scipy.stats.kruskal()) if data are non-normally distributed. The Kruskal-Wallis test is a non-parametric alternative that does not assume normality and is appropriate for count data that may be skewed.

If the omnibus test is statistically significant, Dunn's post-hoc test with Bonferroni correction (scikit_posthocs.posthoc_dunn()) will identify which specific racial/ethnic group pairs differ significantly.

For the binary outcome (excessive diagnostics), a chi-square test (scipy.stats.chi2_contingency()) or Fisher's exact test will compare the proportion of patients receiving excessive diagnostics across racial/ethnic groups.

Model to Create and Test:

To adjust for potential confounders, a multivariable negative binomial regression model will be fit using statsmodels.api.GLM(family=sm.families.NegativeBinomial()) with additional procedures count as the outcome and race/ethnicity as the main predictor. The model will adjust for age category, sex, and hypertension status. This model accounts for overdispersion common in count data. Adjusted incidence rate ratios (IRRs) with 95% confidence intervals will be generated, using Non-Hispanic White as the reference group. The model equation is:

$\log(\text{Additional\_Procedures}) = \beta_0 + \beta_1(\text{Race/Ethnicity}) + \beta_2(\text{Age\_Category}) + \beta_3(\text{Sex}) + \beta_4(\text{Hypertension})$

Tables and Figures:

- Table 5: Descriptive statistics of additional procedures by race/ethnicity (N, mean, SD, median, IQR)
- Table 6: Proportion with excessive diagnostics by race/ethnicity
- Table 7: Adjusted analysis results showing incidence rate ratios from negative binomial regression
- Figure 2: Box plot with swarm overlay showing distribution of additional procedures by race/ethnicity

All analyses will be performed using Python version 3.13.

Objective 4: Figures as part of the SAP are added here:

1. Figure 1: Grouped bar chart - PE incidence rates by age group, stratified by hypertension status (addresses Aims 1 and 2)

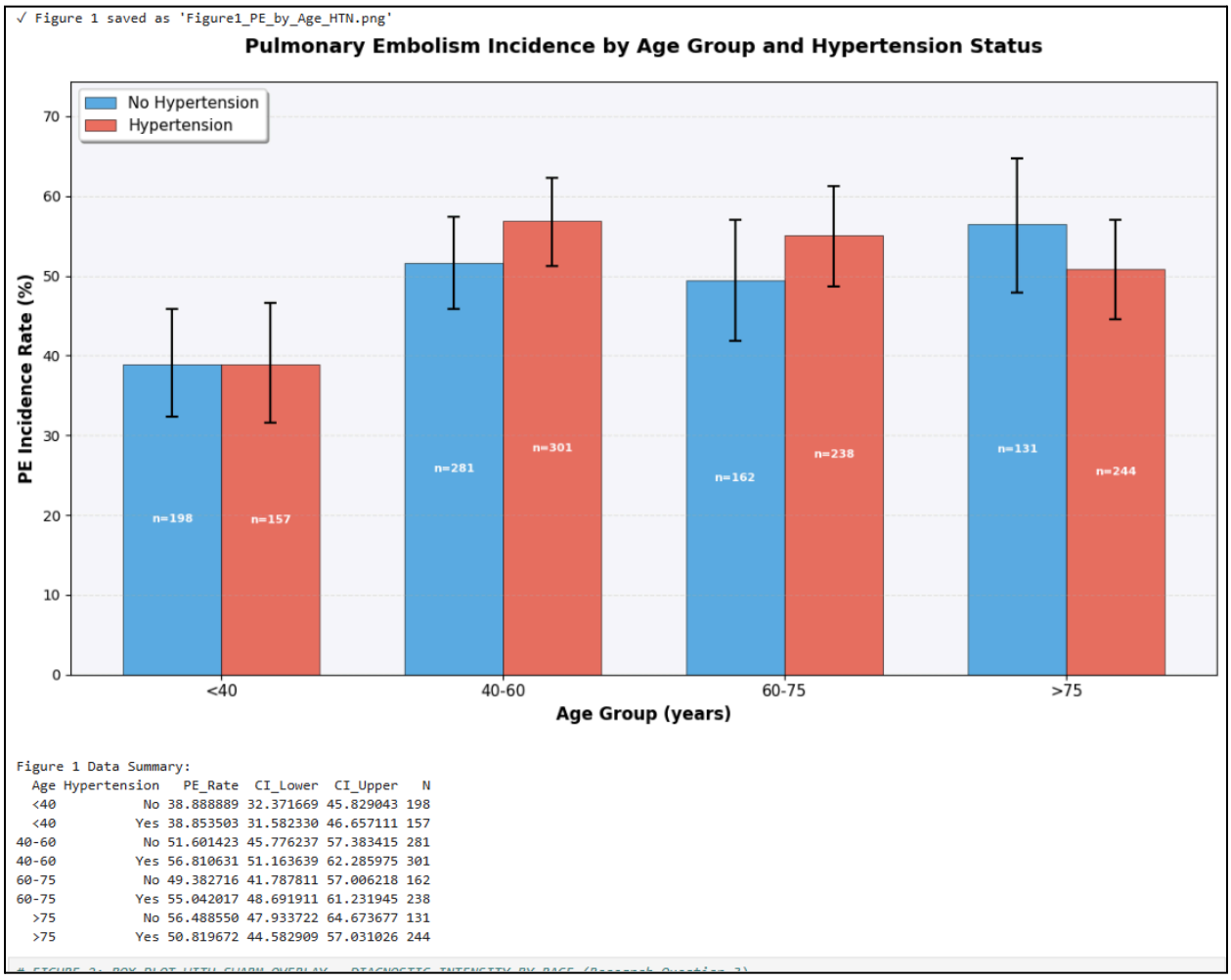

Figure 1 Data Summary:

Age	Hypertension	PE_Rate	CI_Lower	CI_Upper	N
<40	No	38.888889	32.371669	45.829043	198
<40	Yes	38.853503	31.582330	46.657111	157
40-60	No	51.601423	45.776237	57.383415	281
40-60	Yes	56.810631	51.163639	62.285975	301
60-75	No	49.382716	41.787811	57.006218	162
60-75	Yes	55.042017	48.691911	61.231945	238
>75	No	56.488550	47.933722	64.673677	131
>75	Yes	50.819672	44.582909	57.031026	244

2. Figure 2: Box plot with swarm overlay - Additional diagnostic procedures by race/ethnicity among PE+ patients (addresses Aim 3)

```
Analyzing 863 PE+ patients across 5 racial/ethnic groups

Kruskal-Wallis Test: H = 2.393, p = 0.6638
✓ Figure 2 saved as 'Figure2_DiagnosticIntensity_by_Race.png'
```

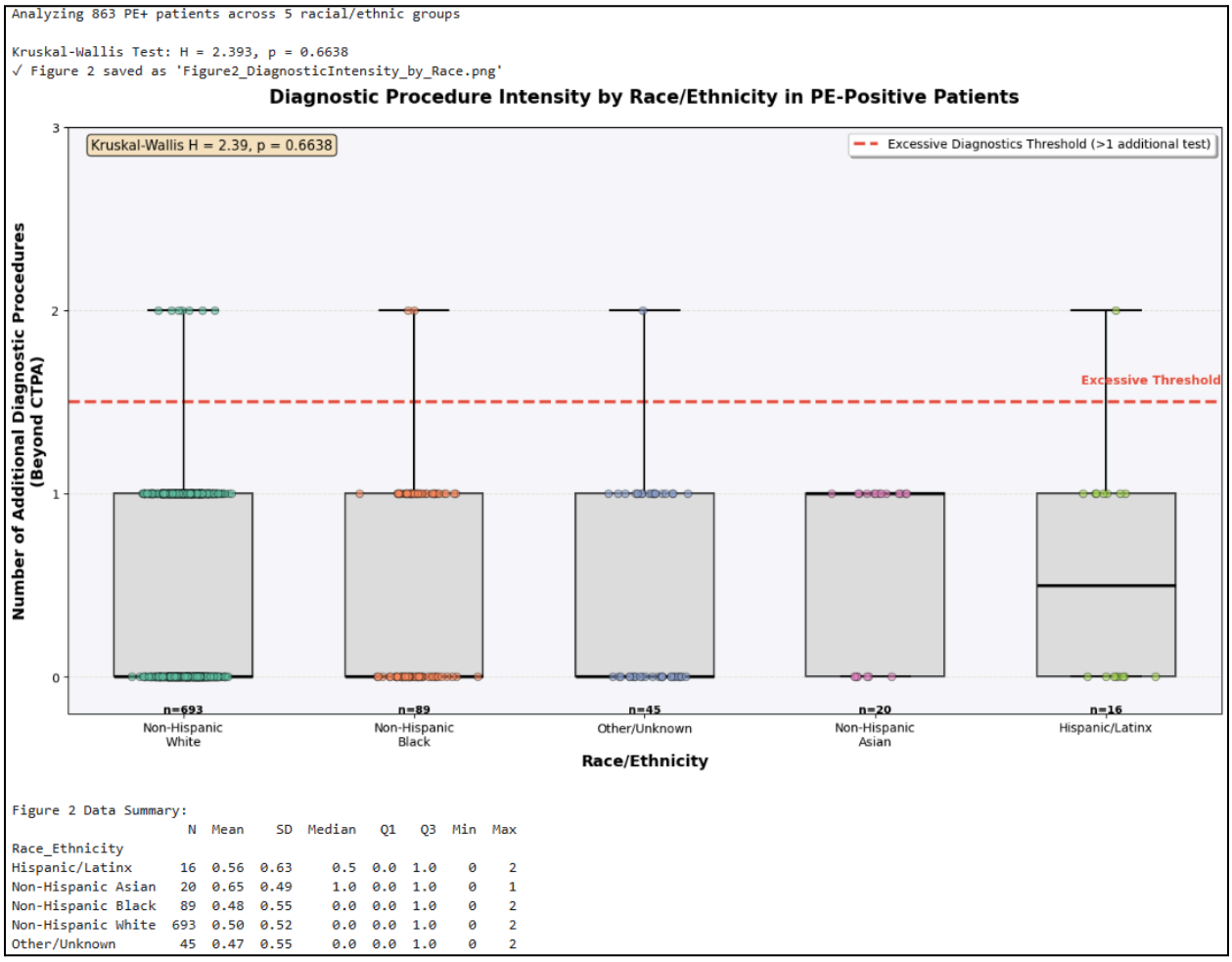

Figure 2 Data Summary:

Race_Ethnicity	N	Mean	SD	Median	Q1	Q3	Min	Max
Hispanic/Latinx	16	0.56	0.63	0.5	0.0	1.0	0	2
Non-Hispanic Asian	20	0.65	0.49	1.0	0.0	1.0	0	1
Non-Hispanic Black	89	0.48	0.55	0.0	0.0	1.0	0	2
Non-Hispanic White	693	0.50	0.52	0.0	0.0	1.0	0	2
Other/Unknown	45	0.47	0.55	0.0	0.0	1.0	0	2

Citations

1. Rashedi, S. (2025). Developing validated tools to identify pulmonary embolism in electronic databases: The PE-EHR+ study (Version V1) [Data set]. Harvard Dataverse. https://doi.org/doi:10.7910/DVN/EQGJU2