Assignment 3 - EHR Data Analysis Results

Authors: Xuange Liang (xl3493), Zexuan Yan (zy2654)

Question 1

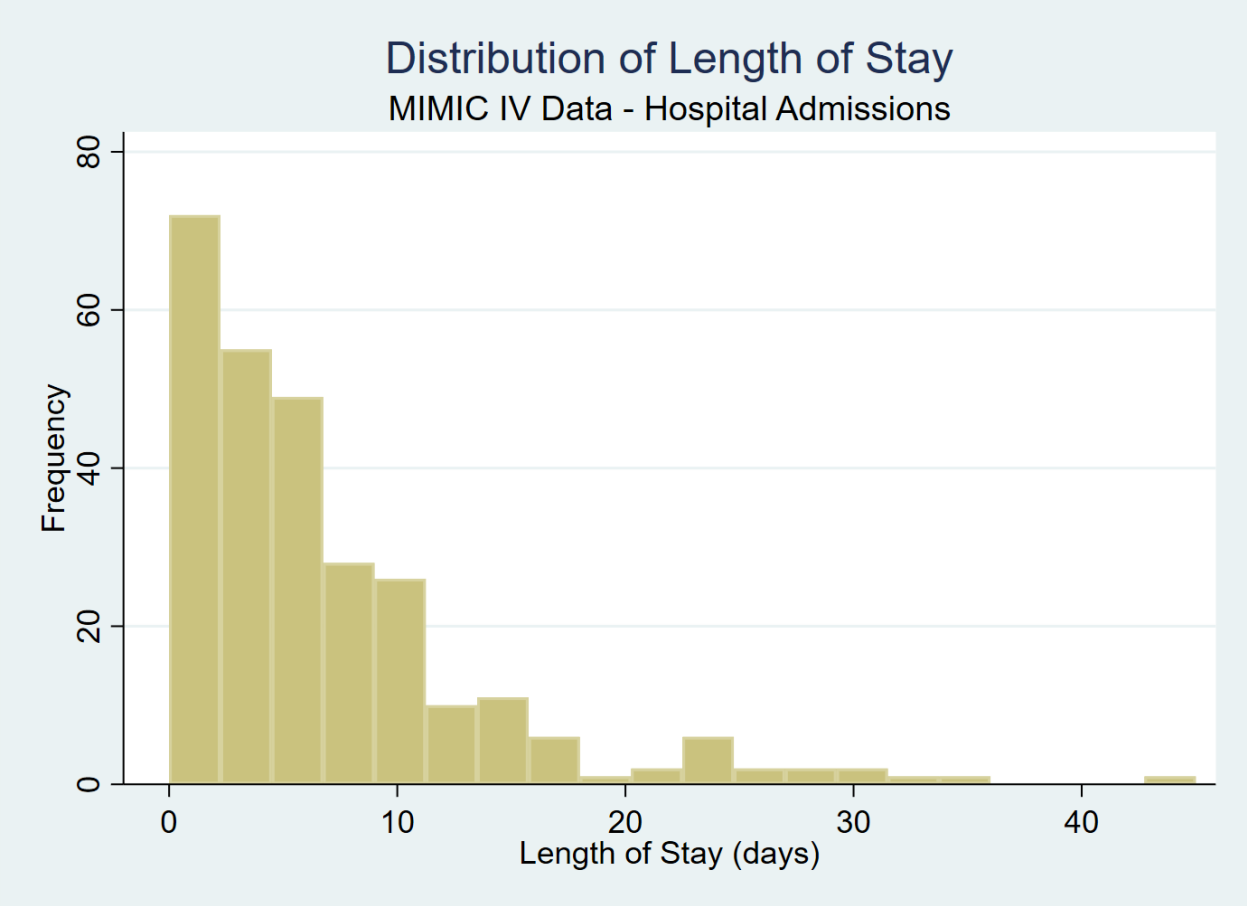

Figure 1: Length of Stay Histogram

Question 2

Septicemia & Disseminated Infections (DRG 720): 4.36%

Acuity Score Statistics: - Mean: 2.26 - Standard Deviation: 0.64

Question 3

Race/Ethnicity Distribution:

Race/Ethnicity	N	Percent
Black	49	17.82%
Hispanic	10	3.64%
Other/Unknown	46	16.73%
White	170	61.82%

Question 4

Imputed Acuity Score Statistics: - Mean: 2.14 - Standard Deviation: 0.58

Question 5

Provider Group P0 has the shortest unadjusted mean length of stay (4.50 days).

Significance Testing vs. P0:

Provider Group	P-value	Significant?
P1	0.1463	No
P2	0.4467	No
P3	0.4370	No
P4	0.0051	Yes*
P5	0.3334	No
P6	0.1871	No
P7	0.2404	No
P8	0.6550	No
P9	0.3555	No

*Indicates significance at =0.05

Only **Provider Group P4** has significantly different LOS compared to P0.

Question 6

Risk-Adjusted Significance Testing vs. P0:

Provider Group	P-value	Significant?
P1	0.0945	No
P2	0.0353	Yes*
P3	0.3395	No
P4	0.0001	Yes*
P5	0.0259	Yes*
P6	0.0335	Yes*
P7	0.0105	Yes*
P8	0.0584	No
P9	0.0245	Yes*

*Indicates significance at =0.05

After risk adjustment, **6 provider groups** (P2, P4, P5, P6, P7, P9) show significant differences from P0, compared to only 1 group (P4) in unadjusted analysis.

Explanation: Risk adjustment eliminates case selection bias by controlling for patient characteristics (age, insurance, DRG, acuity) that affect LOS but are outside providers' control. Some providers care for sicker patients, which masked their efficiency in unadjusted comparisons. Risk adjustment reveals true differences in provider performance by accounting for case mix.

Question 7

Maximum number of diagnoses per person: 170

Question 8

Stata Code for Reshaping:

```
use diagnosis, clear

* Create sequential diagnosis ID for each patient
bysort subject_id: generate dx_id = _n

* Keep essential variables
keep subject_id dx_id icd_code icd_title

* Reshape from long to wide format
reshape wide icd_code icd_title, i(subject_id) j(dx_id)

* Display results
describe icd_code1 icd_code2 icd_code3
list subject_id icd_code1 icd_code2 icd_code3 in 1/5
```

This converts the data from long format (5,051 rows, multiple diagnoses per patient) to wide format (100 rows, one row per patient with diagnoses in columns: icd_code1, icd_code2, etc.).