# Automation of Statistics Summary and Analysis

Lihai Song, MS[1], Zi Wang, MS[1], Jing Huang, PHD,[1,2]
1.Department of Biomedical and informatics, The children's Hospital of Philadelphia
2.Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania Perelman School of Medicine

## BACKGROUND

**Motivation**

Basic statistical analyses, e.g. summary statistics, regression analysis and testing of association, are used in almost all the medical research. The analyses can be repetitive and require tedious work of logging to ensure reproducibility of the study. To make statistics easier for medical searchers, we developed a User Interface (UI) of statistical analysis which automates basic statistical analysis, produces publishable tables and figures with just simple clicks. Our UI will provide a convenient tool for medical researchers to explore data, conduct basic analysis and increase reproducibility.

**The developed user interface**

●Easy to use for researchers with less statistical programming experience.

●Fast response substantially improve research efficiency.

●Reduce errors by hand coding and produce publishable tables and figures.

## OBJECTIVE

●To reduce research burden on researchers by providing automated statistical analysis tools.

●To provide various analysis with simple clicks.

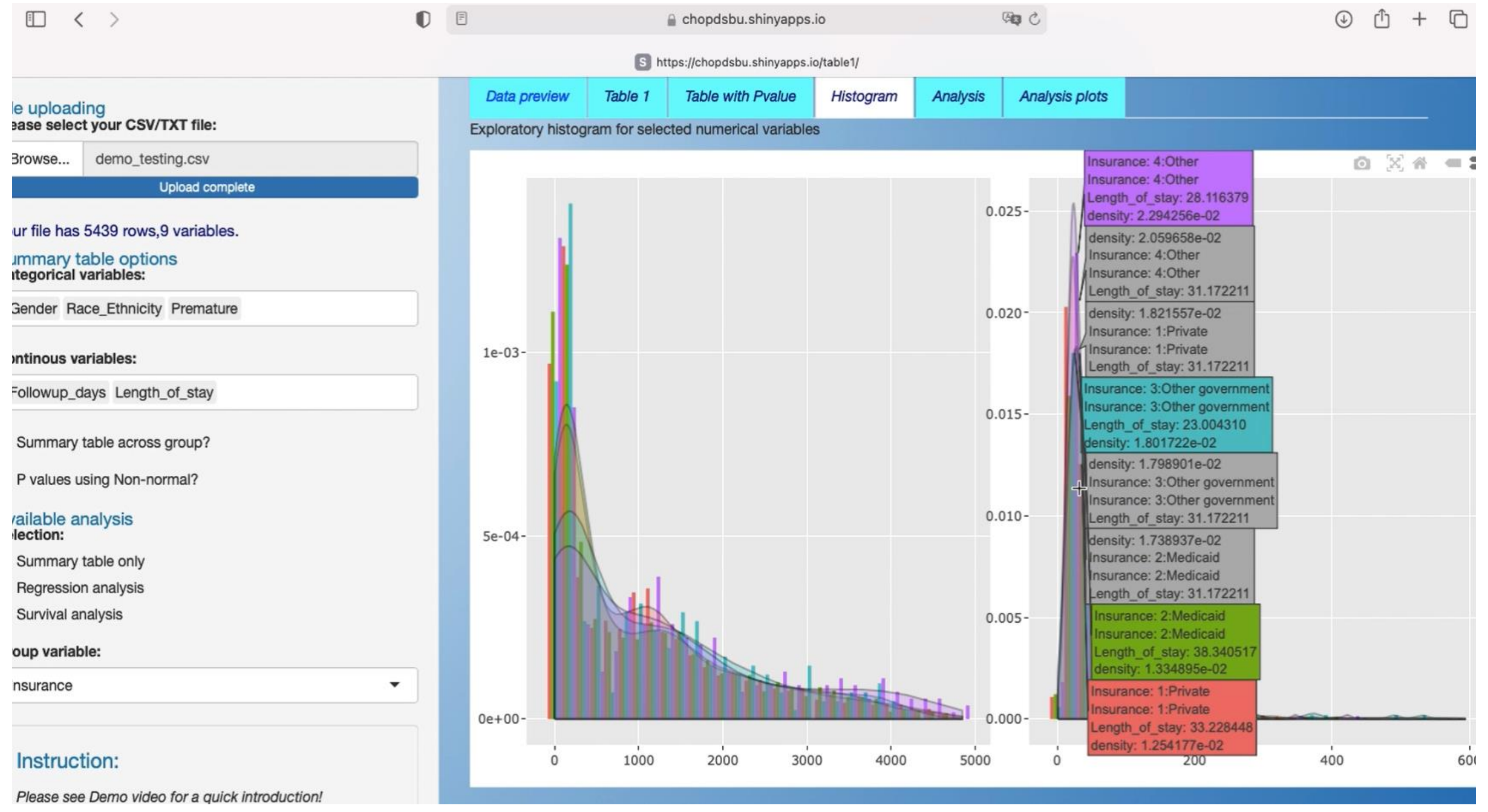
●To provide downloadable and publishable tables and figures.

●To improve research reproducibility.

## FEATURES

**●User friendly interface:**

●This UI provides typing-free click & select interface including a built-in tutorial video and an analysis results with download buttons.



le uploading
ease select your CSV/TXT file:
Browse... demo_testing.csv
Upload complete
ur file has 5439 rows,9 variables.
ummary table options
ategorical variables:
Gender Race_Ethnicity Premature
ontinous variables:
Followup_days Length_of_stay
Summary table across group?
P values using Non-normal?
ailable analysis
lection:
Summary table only
Regression analysis
Survival analysis
oup variable:
nsurance
Instruction:
*Please see Demo video for a quick introduction!*

Exploratory histogram for selected numerical variables

Insurance: 4:Other
Insurance: 4:Other
Length_of_stay: 28.116379
density: 2.294256e-02

density: 2.059658e-02
Insurance: 4:Other
Insurance: 4:Other
Length_of_stay: 31.172211

density: 1.821557e-02
Insurance: 1:Private
Insurance: 1:Private
Length_of_stay: 31.172211

Insurance: 3:Other government
Insurance: 3:Other government
Length_of_stay: 23.004310
density: 1.801722e-02

density: 1.798901e-02
Insurance: 3:Other government
Insurance: 3:Other government
Length_of_stay: 31.172211

density: 1.738937e-02
Insurance: 2:Medicaid
Insurance: 2:Medicaid
Length_of_stay: 31.172211

Insurance: 2:Medicaid
Insurance: 2:Medicaid
Length_of_stay: 38.340517
density: 1.334895e-02

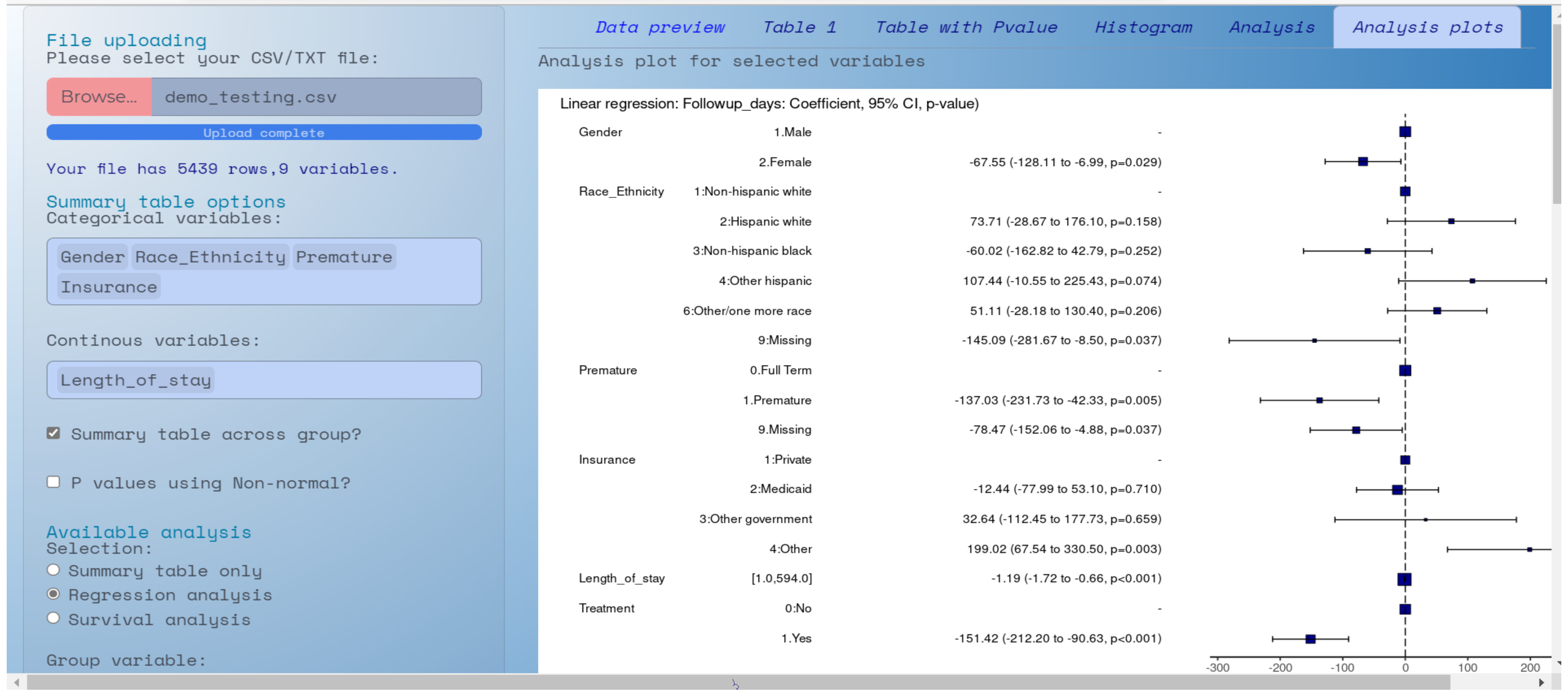
Insurance: 1:Private
Insurance: 1:Private
Length_of_stay: 33.228448
density: 1.254177e-02

**●Instantaneous output**

- ●Demographics and baseline characteristics summary statistics and interactive graph functionality of continuous variables.
- ● Regression analysis, e.g. logistic regression, linear regression, and survival analysis can be performed. Publication-quality tables and figures can be downloaded directly from the UI.

## RESULTS

### Figure 1. Main page of the UI



File uploading
Please select your CSV/TXT file:
Browse... demo_testing.csv
Upload complete

Your file has 5439 rows,9 variables.

Summary table options
Categorical variables:
Gender Race_Ethnicity Premature Insurance

Continous variables:
Length_of_stay

☑ Summary table across group?

☐ P values using Non-normal?

Available analysis
Selection:
○ Summary table only
● Regression analysis
○ Survival analysis

Group variable:

Analysis plot for selected variables

Linear regression: Followup_days: Coefficient, 95% CI, p-value)

| | | |
|---|---|---|
| Gender | 1.Male | - |
| | 2.Female | -67.55 (-128.11 to -6.99, p=0.029) |
| Race_Ethnicity | 1:Non-hispanic white | - |
| | 2:Hispanic white | 73.71 (-28.67 to 176.10, p=0.158) |
| | 3:Non-hispanic black | -60.02 (-162.82 to 42.79, p=0.252) |
| | 4:Other hispanic | 107.44 (-10.55 to 225.43, p=0.074) |
| | 6:Other/one more race | 51.11 (-28.18 to 130.40, p=0.206) |
| | 9:Missing | -145.09 (-281.67 to -8.50, p=0.037) |
| Premature | 0.Full Term | - |
| | 1.Premature | -137.03 (-231.73 to -42.33, p=0.005) |
| | 9.Missing | -78.47 (-152.06 to -4.88, p=0.037) |
| Insurance | 1:Private | - |
| | 2:Medicaid | -12.44 (-77.99 to 53.10, p=0.710) |
| | 3:Other government | 32.64 (-112.45 to 177.73, p=0.659) |
| | 4:Other | 199.02 (67.54 to 330.50, p=0.003) |
| Length_of_stay | [1.0,594.0] | -1.19 (-1.72 to -0.66, p<0.001) |
| Treatment | 0:No | - |
| | 1.Yes | -151.42 (-212.20 to -90.63, p<0.001) |

### Figure 2. Summary Statistics(Table 1)

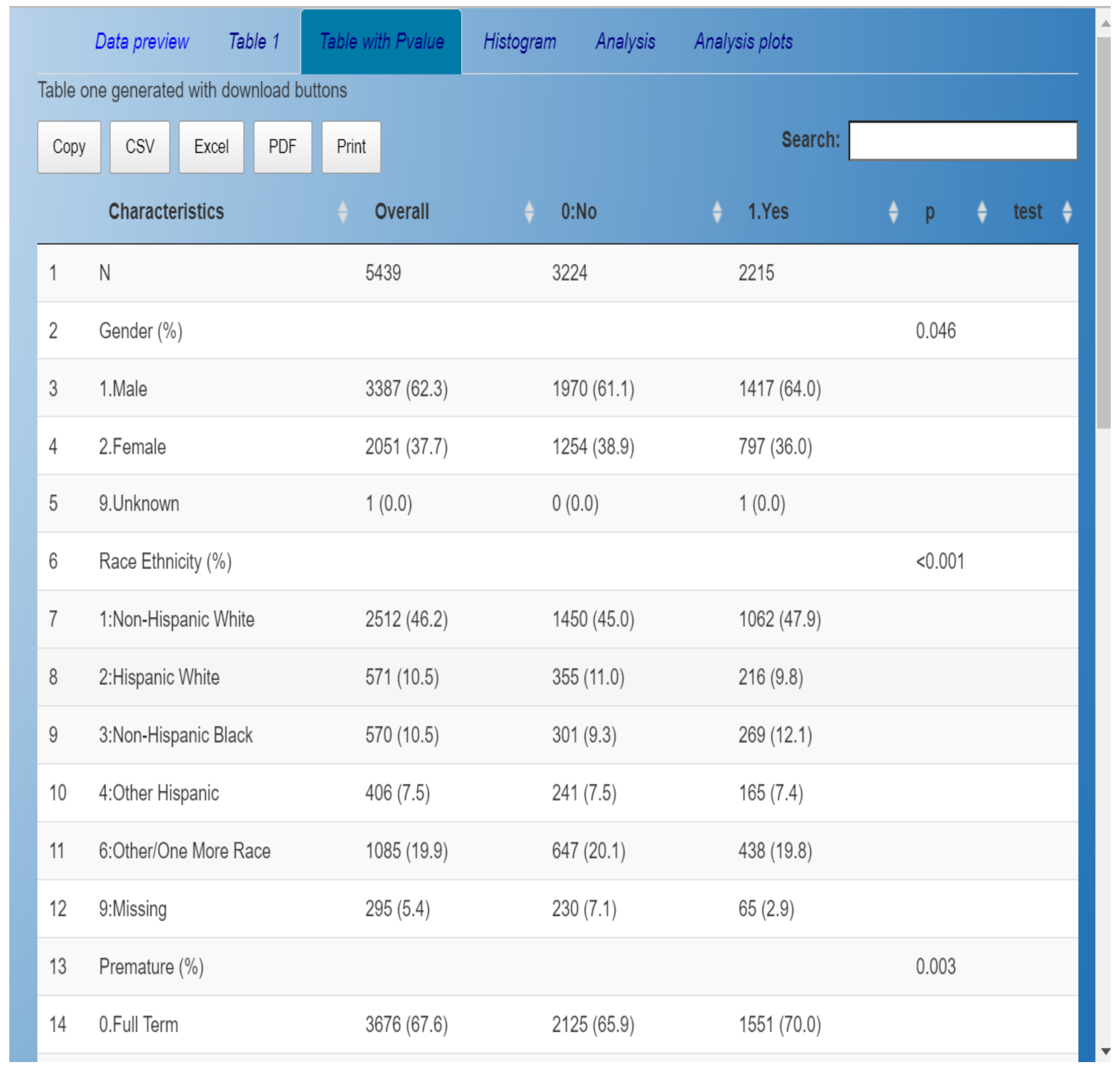
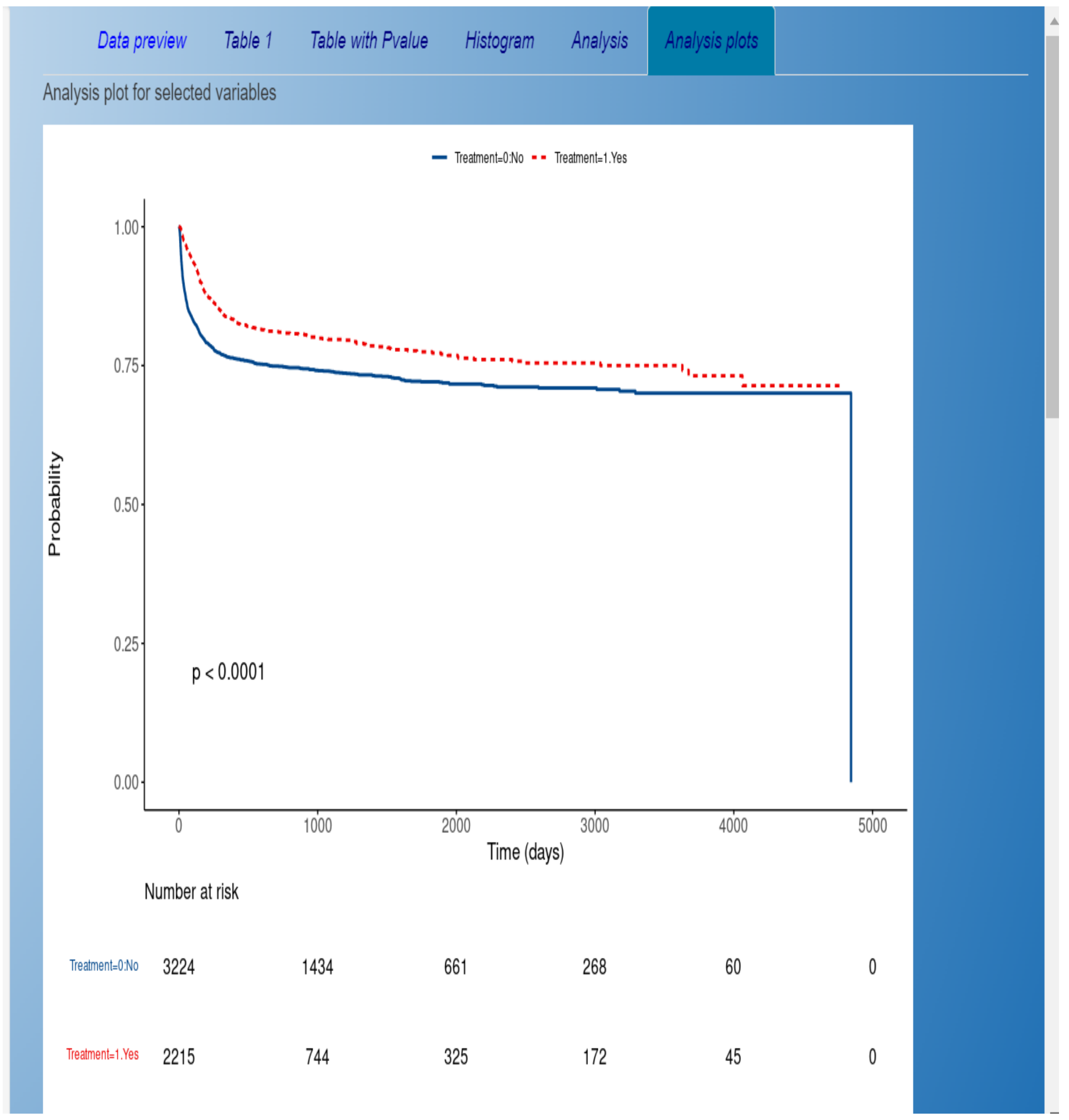


Table one generated with download buttons

| | Characteristics | Overall | 0:No | 1.Yes | p | test |
|---|---|---|---|---|---|---|
| 1 | N | 5439 | 3224 | 2215 | | |
| 2 | Gender (%) | | | | 0.046 | |
| 3 | 1.Male | 3387 (62.3) | 1970 (61.1) | 1417 (64.0) | | |
| 4 | 2.Female | 2051 (37.7) | 1254 (38.9) | 797 (36.0) | | |
| 5 | 9.Unknown | 1 (0.0) | 0 (0.0) | 1 (0.0) | | |
| 6 | Race Ethnicity (%) | | | | <0.001 | |
| 7 | 1:Non-Hispanic White | 2512 (46.2) | 1450 (45.0) | 1062 (47.9) | | |
| 8 | 2:Hispanic White | 571 (10.5) | 355 (11.0) | 216 (9.8) | | |
| 9 | 3:Non-Hispanic Black | 570 (10.5) | 301 (9.3) | 269 (12.1) | | |
| 10 | 4:Other Hispanic | 406 (7.5) | 241 (7.5) | 165 (7.4) | | |
| 11 | 6:Other/One More Race | 1085 (19.9) | 647 (20.1) | 438 (19.8) | | |
| 12 | 9:Missing | 295 (5.4) | 230 (7.1) | 65 (2.9) | | |
| 13 | Premature (%) | | | | 0.003 | |
| 14 | 0.Full Term | 3676 (67.6) | 2125 (65.9) | 1551 (70.0) | | |

### Figure 3. Survival Analysis graph



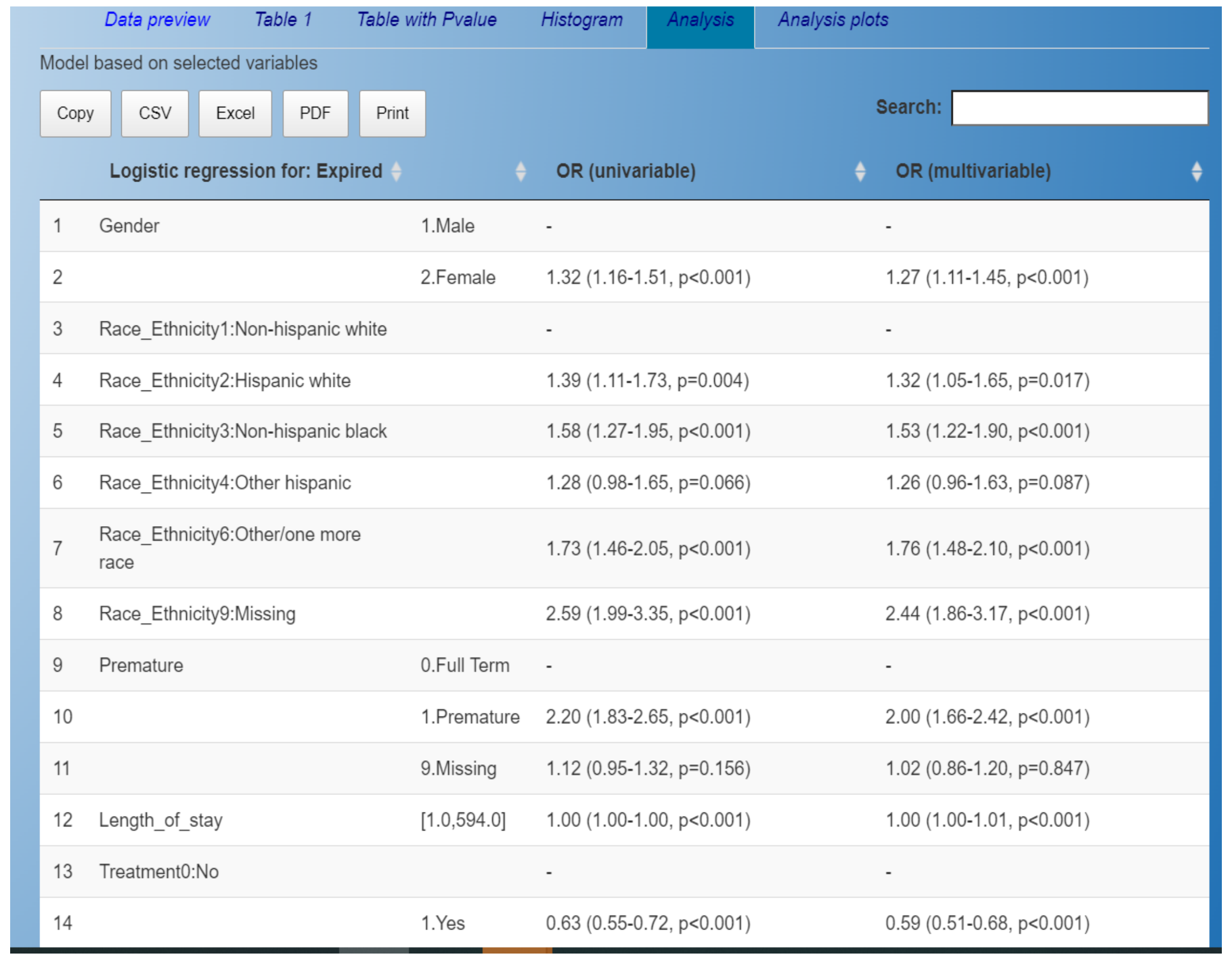
Analysis plot for selected variables

Number at risk

| | | | | | | |
|---|---|---|---|---|---|---|
| Treatment=0:No | 3224 | 1434 | 661 | 268 | 60 | 0 |
| Treatment=1.Yes | 2215 | 744 | 325 | 172 | 45 | 0 |

This research is partially supported by the Children's Hospital of Philadelphia and University of Pennsylvania startup fund, and National Institute of Health (NIH R01HD099348).

### Figure 4: Univariate and Multivariate Logistic Regression Analysis Results



Model based on selected variables

| | Logistic regression for: Expired | | OR (univariable) | OR (multivariable) |
|---|---|---|---|---|
| 1 | Gender | 1.Male | - | - |
| 2 | | 2.Female | 1.32 (1.16-1.51, p<0.001) | 1.27 (1.11-1.45, p<0.001) |
| 3 | Race_Ethnicity1:Non-hispanic white | | - | - |
| 4 | Race_Ethnicity2:Hispanic white | | 1.39 (1.11-1.73, p=0.004) | 1.32 (1.05-1.65, p=0.017) |
| 5 | Race_Ethnicity3:Non-hispanic black | | 1.58 (1.27-1.95, p<0.001) | 1.53 (1.22-1.90, p<0.001) |
| 6 | Race_Ethnicity4:Other hispanic | | 1.28 (0.98-1.65, p=0.066) | 1.26 (0.96-1.63, p=0.087) |
| 7 | Race_Ethnicity6:Other/one more race | | 1.73 (1.46-2.05, p<0.001) | 1.76 (1.48-2.10, p<0.001) |
| 8 | Race_Ethnicity9:Missing | | 2.59 (1.99-3.35, p<0.001) | 2.44 (1.86-3.17, p<0.001) |
| 9 | Premature | 0.Full Term | - | - |
| 10 | | 1.Premature | 2.20 (1.83-2.65, p<0.001) | 2.00 (1.66-2.42, p<0.001) |
| 11 | | 9.Missing | 1.12 (0.95-1.32, p=0.156) | 1.02 (0.86-1.20, p=0.847) |
| 12 | Length_of_stay | [1.0,594.0] | 1.00 (1.00-1.00, p<0.001) | 1.00 (1.00-1.01, p<0.001) |
| 13 | Treatment0:No | | - | - |
| 14 | | 1.Yes | 0.63 (0.55-0.72, p<0.001) | 0.59 (0.51-0.68, p<0.001) |

## LIMITATIONS

- ● Subgroup is not supported in summary table.
- ● Interaction terms are not included in all analysis reports.
- ●The accepted CSV file has size limit.
- ●This UI can NOT do data cleaning and it requires high quality of analytic data set.

## CONCLUSIONS

●The developed UI automates basic statistical analysis, substantially reduces errors of hand coding, and improves the productivity of research.

●The developed UI makes data exploration and analysis easy for medical researchers, it increases the efficiency and productivity of medical research.

●Analysis using advanced statistical methods are under-development.

## FUTURE WORK

**●More to come:**

●We are working on other commonly used basic statistical analysis.

●Advanced statistical analysis methods will be included in the next version.

●More file formats will be supported in the next version.

● Summary table will allow subgroup statistics.