Udacity, Nanodegree: Data Scientist

Capstone Project:

**Maternal Health** - Maternal Sepsis by Select Risk Factors

By Laura Hagg

Table of contents

[1. The Dataset 1](#_Toc136462935)

[2. Data Cleaning 2](#_Toc136462936)

[2.1 Changing the datatype of the p-value to a float 2](#_Toc136462937)

[2.2 Removing duplicated information regarding Tobacco and Alcohol use 2](#_Toc136462938)

[3. Questions 3](#_Toc136462939)

[3.1 What are the ten most risk factors to get a sepsis during the maternal window? 3](#_Toc136462940)

[3.2 From this top ten, which are statistically significant (p-value < 0.05)? 4](#_Toc136462941)

[3.2.1 During Pregnancy 5](#_Toc136462942)

[3.2.2 During Delivery 5](#_Toc136462943)

[3.3.3 Postpartum 6](#_Toc136462944)

[3.3 What are the top ten risk factors with the highest incidence for a statistically significant sepsis (any or severe)? 7](#_Toc136462945)

[3.3.1 Pregnancy 7](#_Toc136462946)

[3.3.2 Delivery 8](#_Toc136462947)

[3.3.3 Postpartum 8](#_Toc136462948)

[3.4 9](#_Toc136462949)

[3.x 9](#_Toc136462950)

[4. Conclusion 9](#_Toc136462951)

# The Dataset

The dataset contains information on the occurrence of maternal sepsis among live births during the pregnancy, delivery, and postpartum periods from 2016 to 2018. It provides counts, rates, and measures of association related to specific risk factors and the incidence of maternal sepsis, identified through administrative means.

The dataset is available on the HealthData.gov website. You can access the dataset using the following link: [Link to the dataset](https://healthdata.gov/State/Maternal-Sepsis-by-Select-Risk-Factors-SPARCS-2016/gy4j-ef7j).

Additional information about this dataset can be found [here](https://health.data.ny.gov/Health/Maternal-Sepsis-by-Select-Risk-Factors-SPARCS-2016/p9ay-x62y). The provided link includes details about the columns and their data types. Additionally, a data dictionary with column descriptions is available in the repository under the filename "HDNY\_MaternalSepsis\_RiskFactors\_DataDictionary\_v2.pdf".

The dataset consists of 16 columns and 585 rows.

# Data Cleaning

The dataset is generally clean, but there were a few tasks I performed before working with the data

## Changing the datatype of the p-value to a float

The issue arises from the columns 'Any\_Sepsis\_p-value' and 'Severe\_Sepsis\_p-value', where non-floating point numbers such as '<0.0001' are present. To facilitate better analysis, I converted these values to '0.0001'.

## 2.2 Removing duplicated information regarding Tobacco and Alcohol use

Due to data collection from different sources ('SPARCS', 'Birth certificate', 'SPARCS & Birth certificate'), redundant information regarding tobacco and alcohol use exists. To ensure data integrity in our analysis, I decided to exclude the risk factors collected solely from the birth certificate, thus eliminating duplicated information.

In the dataframe, there are two risk factors that appear twice:

1. Tobacco use (& Tobacco Use)
2. Alcohol use (& Alcohol abuse)

The reason for this duplication is that the data is collected from two different sources: SPARCS and Birth Certificate.

To address this issue, I have three options:

1. Keep the data as it is.
2. Combine the data together.
3. Remove one of the risk factors.

After careful consideration, I have decided to remove one of these risk factors to avoid duplication in the data analysis. Specifically, I will remove the risk factors obtained from the 'Birth Certificate' data source. This decision is based on the fact that the 'Tobacco use' and 'Alcohol use' risk factors from the 'SPARCS' source have a simpler structure, with two options (yes and no), while the risk factors from the 'Birth Certificate' source have three options (yes, no, and unknown). By removing the risk factors from the 'Birth Certificate' source, the analysis will be easier and more consistent when exploring tobacco and alcohol use.

# Questions

## What are the ten most risk factors to get a sepsis during the maternal window?

With this question I want to show the ten most often risk factors during the three maternal windows (pregnancy, delivery and postpartum). Top ten means the highest number of livebirths with women who have these risk factors. For the first question I did not take into account if the sepsis is ‘any sepsis or ‘severe sepsis’.

|  |  |
| --- | --- |
| During pregnancy:   * 'Nulliparity', * 'Group B Strep Carrier', * 'Previous cesarean delivery', * 'Obesity', * 'Blood loss anemia', * 'Gestational diabetes mellitus', * 'Deficiency anemia', * 'Chronic pulmonary disease', * 'Asthma', * 'Gestational hypertension' |  |
| During delivery:   * 'Induction of Labor', * 'Cesarean Delivery', * 'Nulliparity', * 'Group B Strep Carrier', * 'Previous cesarean delivery', * 'Obesity', * 'Blood loss anemia', * 'Premature rupture of membranes', * 'Gestational diabetes mellitus', * 'Episiotomy' |  |
| Postpartum:   * 'Induction of Labor', * 'Cesarean Delivery', * 'Nulliparity', * 'Group B Strep Carrier', * 'Previous cesarean delivery', * 'Obesity', * 'Blood loss anaemia', * 'Premature rupture of membranes', * 'Gestational diabetes mellitus', * 'Episiotomy' |  |

As you can see during delivery and postpartum there are the same top ten risk factors.

## From this top ten, which are statistically significant (p-value < 0.05)?

The first question just answered how many women had each risk factor, but it did not determine if there is a statistically significant risk of developing sepsis. The next visuals will show which of these risk factors have a p-value < 0.05 for both any sepsis and severe sepsis.

### During Pregnancy

|  |  |
| --- | --- |
| Any Sepsis | Severe Sepsis |
|  |  |

During pregnancy, there is a significant risk for any sepsis and severe sepsis if women have risk factors like: Nulliparity, Gestational diabetes mellitus, Deficiency anemia, and chronic pulmonary disease. For the risk factor obesity, it is only a significant risk factor for any sepsis, but not for severe sepsis.

### 3.2.2 During Delivery

|  |  |
| --- | --- |
| Any Sepsis | Severe Sepsis |
|  |  |

During delivery there is a significant risk for both, any and severe, sepsis, if women have the following risk factors:

* 'Cesarean Delivery',
* 'Group B Strep Carrier',
* 'Obesity',
* 'Blood loss anaemia',

Induction of Labor, Previous cecarean delivery, Gestational diabetes mellitus and Episiotomy have a significant risk for any sepsis. And a premature rupture of membranes leads to a significant risk for a severe sepsis.

## 3.3.3 Postpartum

|  |  |
| --- | --- |
| Any Sepsis | Severe Sepsis |
|  |  |

## What are the top ten risk factors with the highest incidence for a statistically significant sepsis (any or severe)?

### 3.3.1 Pregnancy

|  |  |
| --- | --- |
|  |  |

### 3.3.2 Delivery

|  |  |
| --- | --- |
|  |  |

### 3.3.3 Postpartum

|  |  |
| --- | --- |
|  |  |

## 3.4

## 3.x

# Conclusion