Eula Fullerton

Professor Williams

DSC540-Data Preparation

Sept 22, 2024

**Milestone 1**

**Project Subject Area:**

I will investigate the relationship between FDA-regulated product usage and natality outcomes in my project. Unfortunately, drug names are not present in the natality data. Therefore, this project will explore how various maternal risk factors, procedures during labor, and health conditions serve as indirect indicators of drug use during pregnancy and how these may correlate with birth outcomes.

**Data Sources:**

1. Flat File such as CSV/Excel/PDF: CDC Natality Data
   1. Description of data source: The CDC Natality Data dataset includes birth outcomes and maternal health information. It includes father characteristics, history and prenatal characteristics, maternal risk factors, pregnancy risk factors, maternal infections present or treated, labor characteristics, delivery characteristics, maternal morbidity, infant characteristics, abnormal conditions of newborn, congenital abnormalities, and causes of infant death.
   2. Link or Flat File uploaded: [Linked Birth / Infant Death Records, 2017-2022 Expanded Request (cdc.gov)](https://wonder.cdc.gov/lbd-current-expanded.html)
2. API: OpenFDA API
   1. Description of data source: The Open Drug Label API provides detailed information about drugs, including their pregnancy categories (A, B, C, D, X)
   2. Link: [openFDA](https://open.fda.gov/)
3. Website:
   1. Description of data source:
      1. Provides three columns
         1. Pharmaceutical agents, Australia, and the United States
      2. Twenty-one columns
   2. Link: [https://en.wikipedia.org/wiki/Pregnancy\_category](%20https://en.wikipedia.org/wiki/Pregnancy_category)

**Relationships**

1. The key relationship in this study will be through drug names prescribed or associated with pregnancy outcomes.

**Project Plan**

1. **Project approach/plan**
   1. It is common for pregnant women to require at least one medicine during their pregnancy. Although many are safe, there is a lack of information regarding specific medications. It is often recommended to always talk with your health care provider before taking any medications, dietary supplements, or even vitamins while pregnant. My goal is to analyze pregnancy-related data to understand the risks associated with drug use during pregnancy. My first step will be producing five questions I want to answer.
      1. What are the most prescribed or non-prescribed drugs during pregnancy?
         1. Requires analyzing CDC data for mentions of medications or inferring drug use based on health conditions.
         2. Querry OpenFDA API for pregnancy-related drugs
      2. How does the use of these drugs correlate with adverse maternal outcomes?
      3. How does drug use during pregnancy impact birth time, birth weight, or congenital anomalies?
      4. Do maternal factors (pre-existing conditions) influence the likelihood?
      5. Which drugs are commonly listed in the Drug Label API that are often associated with pregnancies?
   2. I plan to download the CDC Natality Data file, access the API, and fetch the Drug Label API to provide pregnancy categories.
   3. Since I am using common drug conditions, I will need to find the drugs commonly prescribed during pregnancy; examples would be drugs for gestational diabetes, pre-eclampsia, or labor induction. This will require a new column referencing the drug typically used to treat the condition.
2. **What concerns/challenges you think you will face with the data/project topic**
   1. I think I will face data quality and consistency, scalability, data integration issues, lack of documentation, or file size limitations.
   2. Another is dealing with uncertainty. In this project, I will be working with indirect data (conditions instead of drug names); therefore, it is essential to acknowledge the assumptions and limitations of the project.
3. **Ethical Implications of your project topic**
   1. This project involves analyzing pregnancy and health data, which requires careful consideration of ethical issues. It is crucial to represent the data responsibly. Evidence and direct medical claims without adequate evidence and acknowledge that a causal relationship between a product and reported reactions cannot be established based on openFDA data alone. Additionally, CDC data use restrictions include using the data solely for health statistical reporting and analysis, avoiding the presentation or publication of birth counts for fewer than ten births, not attempting to identify individuals or establishments in the data, and not disclosing or using any personal or establishment identity inadvertently discovered.
   2. Additionally, it is important not to create assumptions that certain conditions involve certain medications and acknowledge the limitations that some medicines may not always be used for the associated conditions and that some conditions may be underreported.

Data Notes:

'Suppressed' appears instead of death, birth, and rate values when the death value represents 1-9 sub-national events. This is because the data do not meet the criteria for confidentiality constraints. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Assurance%20of%20Confidentiality)

'Unreliable' appears for rates when there are fewer than 20 deaths in the numerator. This is because the figure does not meet the NCHS standard of reliability or precision. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Unreliable)

Infant deaths are weighted so numbers may not exactly add to totals due to rounding. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#WeightFactors-Totals)

Note that there may be slight differences in the number of infant deaths when comparing the Linked Birth / Death Records to the other vital statistics. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Weight%20Factors)

|  |  |
| --- | --- |
| **Suggested Citation:** | |
| Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Linked Birth / Infant Deaths on CDC WONDER Online Database. Data are from the Linked Birth / Infant Deaths Records 2017-2022, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/lbd-current-expanded.html on Sep 30, 2024 9:02:43 AM  Request can not return more than 75,000 rows  'Suppressed' appears instead of death, birth, and rate values when the death value represents 1-9 sub-national events. This is because the data do not meet the criteria for confidentiality constraints. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Assurance%20of%20Confidentiality)  Infant deaths are weighted so numbers may not exactly add to totals due to rounding. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#WeightFactors-Totals)  Note that there may be slight differences in the number of infant deaths when comparing the Linked Birth / Death Records to the other vital statistics. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Weight%20Factors)  Changes to cause of death classification affect reporting trends. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#ICD-10%20Changes)  Will just focus on pre-pregnancy diabetes/gestational diabetes as well as census region   * **Census Regions and Divisions from the United States Census Bureau:**   Census Region 1 - Northeast:   * + Division 1 - New England:    Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont.   + Division 2 - Middle Atlantic:    New Jersey, New York, Pennsylvania.   Census Region 2 - Midwest:   * + Division 3 - East North Central:    Illinois, Indiana, Michigan, Ohio, Wisconsin.   + Division 4 - West North Central:    Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota.   Census Region 3 - South:   * + Division 5 - South Atlantic:    Delaware, District of Columbia reporting area, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia.   + Division 6 - East South Central:    Alabama, Kentucky, Mississippi, Tennessee.   + Division 7 - West South Central:    Arkansas, Louisiana, Oklahoma, Texas.   Census Region 4 - West:   * + Division 8 - Mountain:    Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming.   + Division 9 - Pacific:    Alaska, California, Hawaii, Oregon, Washington.   Year of Death 2022.  'Suppressed' appears instead of death, birth, and rate values when the death value represents 1-9 sub-national events. This is because the data do not meet the criteria for confidentiality constraints. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Assurance%20of%20Confidentiality)  'Unreliable' appears for rates when there are fewer than 20 deaths in the numerator. This is because the figure does not meet the NCHS standard of reliability or precision. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Unreliable)  Infant deaths are weighted so numbers may not exactly add to totals due to rounding. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#WeightFactors-Totals)  Note that there may be slight differences in the number of infant deaths when comparing the Linked Birth / Death Records to the other vital statistics. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Weight%20Factors)  Changes to cause of death classification affect reporting trends. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#ICD-10%20Changes)  Eclampsia data are coded to "Not Reported" for births to mothers residing in the following reporting areas and years: South Carolina 2017-2021; Tennessee 2017-2020. [More information.](https://wonder.cdc.gov/wonder/help/lbd-expanded.html#Eclampsia) |  |

Adjustment created to reflect sensitive information privacy.

Focus on these features:

* + 1. Maternal Age
    2. Gestational Age
    3. Birth Weight
    4. Cause of infant death
    5. Race/Ethnicity