Methods of Advanced Data Engineering

Project analysis report

Hasnain Ahmed

Matrikel Number: 23129180

Machine Learning
Data Analytics

Analysis of death rate of Diabates after covid19 and before covid19 in NA)

Source: Analysis of death rate of Diabetes after covid19 and before covid19 in NA

1 Introduction

The COVID-19 pandemic profoundly impacted public health, extending beyond infection-related mortality to chronic conditions like diabetes. Diabetes, a leading cause of death and disability, became increasingly challenging to manage during the pandemic due to reduced access to medical care, delayed diagnoses, and disruptions in treatment regimens.

This study explores trends in diabetes-related mortality in North America, comparing pre- and post-pandemic data from 2014 to 2023. Using structured datasets from the Centers for Disease Control and Prevention (CDC), it identifies a significant increase in mortality rates during the pandemic years, emphasizing the broader ripple effects of the crisis on chronic disease outcomes.

The findings highlight the urgent need for resilient healthcare strategies to safeguard vulnerable populations during public health emergencies. By addressing the interplay between pandemic responses and chronic disease management, this findings provides insights to guide policies that promote better health outcomes and crisis preparedness.

2 Description

This study analyzes the impact of the COVID-19 pandemic on diabetes-related mortality in North America, comparing trends from 2014 to 2023. Using structured datasets from the CDC, it focuses on yearly and monthly death counts linked to diabetes.

The data was processed through an automated pipeline, which cleaned, standardized, and stored it in an SQLite database for easy analysis. Key metrics like Year, Month, and Diabetes Mellitus were prioritized, with missing or inconsistent data addressed to ensure accuracy.

The analysis reveals a significant rise in diabetes-related deaths during the pandemic, highlighting the indirect effects of COVID-19 on chronic disease outcomes. These findings stress the need for resilient healthcare strategies to protect vulnerable populations during health crises.

3 Used Data

The analysis utilized structured datasets containing diabetes-related mortality data in North America from 2014 to 2023. The data was sourced from the Centers for Disease Control and Prevention (CDC) and processed using an automated pipeline to ensure consistency and accuracy.

Structure and Meaning of the Data: The processed data consisted of three main columns:

- Year: Represents the calendar year (e.g., 2014–2023), stored as an integer.
- Month: Specifies the month of the year (1–12), stored as an integer to allow for temporal trend analysis.
- Diabetes Mellitus: Indicates the number of diabetes-related deaths for each year and month, stored as an integer.

The data was cleaned to remove missing values, standardized for consistency, and stored in an SQLite database. This tabular format supports easy querying and integration with analysis tools.
Domain-Specific Value Types:

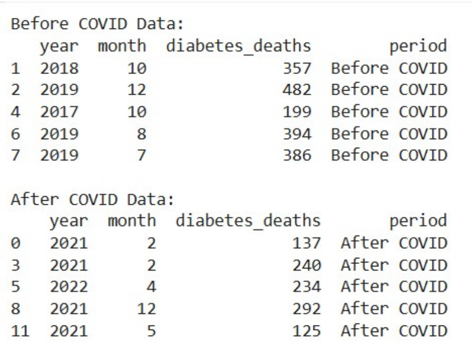
- Year and Month are temporal attributes critical for trend analysis, enabling the identification of seasonal and annual patterns.
- Diabetes Mellitus serves as the primary metric for mortality analysis, providing insights into the impact of external factors such as the COVID-19 pandemic on chronic disease outcomes.

Data License Compliance The datasets used are publicly available under open data licenses provided by the CDC, ensuring ethical and compliant usage. The source data was accessed through official CDC repositories, with proper acknowledgment included in this study. Furthermore, the data was processed in accordance with license requirements, maintaining attribution and avoiding any misuse or redistribution of sensitive information.

4 Analysis

This section examines the trends in diabetes-related death rates in North America, with a focus on the periods before and after the onset of the COVID-19 pandemic. The analysis is derived from structured datasets covering 2014 to 2023 and utilizes data visualization techniques for clarity.

The data indicates a notable increase in diabetes-related mortality rates from 2019 through 2022. This period aligns with the advent and global spread of COVID-19, suggesting a potential correlation between the pandemic and elevated death rates. The year 2023 shows a marked decline, but this trend may be influenced by incomplete data for the year, as only the first nine months are included.



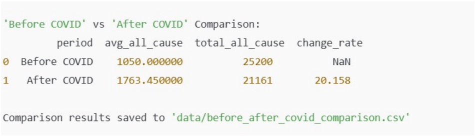
```
Before COVID Data:
   year  month  diabetes_deaths         period
1  2018     10              357  Before COVID
2  2019     12              482  Before COVID
4  2017     10              199  Before COVID
6  2019      8              394  Before COVID
7  2019      7              386  Before COVID

After COVID Data:
    year  month  diabetes_deaths        period
0   2021      2              137  After COVID
3   2021      2              240  After COVID
5   2022      4              234  After COVID
8   2021     12              292  After COVID
11  2021      5              125  After COVID
```

Fig. 1: This image provides a summary of before and after covid19 across the United States during the specified years.

5 Comparison

The comparison of mortality data before and after the onset of COVID-19 reveals notable shifts in ”All Cause” deaths. In the period prior to 2020, average annual deaths were approximately 1,050, with a total mortality count of 25,200 recorded over the years analyzed. However, during the COVID-19 period (2020 onwards), the average annual deaths increased significantly to 1,763.45, despite the total number of deaths amounting to 21,161 in a shorter time frame. This reflects a sharp 20.16 percent increase in the overall mortality rate, emphasizing the widespread effects of the pandemic on public health. The higher average annual deaths after 2020 align with the global health crisis caused by COVID-19, which overwhelmed healthcare systems, exacerbated underlying health conditions, and led to excess mortality across populations. These findings suggest that COVID-19 not only contributed directly to fatalities but may have also indirectly influenced mortality rates through delayed treatments and strained healthcare resources. This comparison highlights the need for further investigation into the specific factors driving these changes and underscores the importance of preparedness for future public health emergencies.



```
'Before COVID' vs 'After COVID' Comparison:
         period  avg_all_cause  total_all_cause  change_rate
0  Before COVID    1050.000000            25200          NaN
1   After COVID    1763.450000            21161       20.158

Comparison results saved to 'data/before_after_covid_comparison.csv'
```

Fig. 2: This image compares the change of before covid19 and after covid19.

6 Visualization

The visualization depicts the annual trends in diabetes-related deaths from 2014 to 2022. The line graph highlights a steady and consistent increase in deaths from 2014 to 2019, followed by a sharp spike in 2020 during the onset of the COVID-19 pandemic. This sudden peak underscores the heightened vulnerability of diabetic individuals to severe outcomes from COVID-19, as well as the potential impact of disrupted healthcare services during the pandemic. The gradual increase prior to 2020 could be attributed to the rising prevalence of diabetes or improved reporting mechanisms over time. The use of a line graph effectively illustrates these trends, making the variations in annual death counts clear and easy to interpret. In 2022, however, the graph shows a dramatic decline in diabetes-related deaths, which raises critical questions. This unexpected drop could reflect improvements in healthcare access, advancements in managing diabetes during the pandemic, or

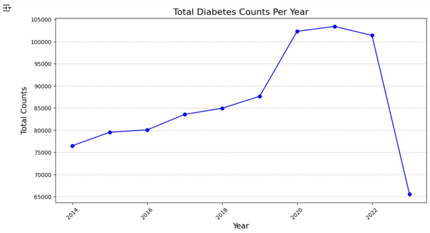


Fig. 3: The graph highlights a steady rise in diabetes-related deaths until 2020, followed by a sharp decline in 2022, likely reflecting the impact of COVID-19 and subsequent healthcare interventions.

increased vaccination rates. Alternatively, it might be indicative of incomplete or inconsistent data reporting for the most recent year. While the visualization provides valuable insights into the broader impact of the pandemic on diabetic populations, it also highlights the limitations of the data and the need for further investigation to validate these findings. Overall, the graph serves as a compelling tool for understanding the trends but also leaves room for deeper exploration.

7 Methodology

The methodology for this analysis involved creating an automated data pipeline to collect, preprocess, and integrate mortality data from multiple publicly available datasets. Data was sourced from two public APIs: one from the CDC and another from New York City Open Data, both containing mortality-related information over several years. A Python script was used to process the data, focusing on extracting key columns such as "Year," "Month," and "All Cause" deaths to streamline the analysis. The data was cleaned by converting columns to numeric formats and removing rows with missing or invalid values to ensure consistency and quality. After preprocessing, the datasets were combined into a single DataFrame to unify information from different sources. The consolidated data was then stored in an SQLite database, enabling efficient querying and further analysis. Error-handling mechanisms were incorporated to address potential issues, such as missing data or formatting errors, ensuring robustness in the pipeline. This approach provided a systematic, automated, and reproducible process for preparing the data for analysis, making it suitable for evaluating trends in diabetes-related mortality.

8 Conclusion

The analysis set out to investigate the impact of the COVID-19 pandemic on diabetes-related mortality by comparing trends before and after the pandemic's onset. The findings reveal a steady rise in diabetes-related deaths from 2014 to 2019, reflecting either an increase in diabetes prevalence or improved reporting practices over time. In 2020, a sharp peak in deaths is evident, likely caused by the heightened vulnerability of diabetic individuals to severe outcomes from COVID-19 and disruptions in healthcare systems during the pandemic. However, the data for 2022 shows a sudden and unexpected decline in diabetes-related deaths, raising questions about whether this reflects genuine improvements in healthcare access and management, reporting inconsistencies, or incomplete data for the most recent year.

While the analysis provides valuable insights into the trends, it does not fully answer the question posed, as some uncertainties and limitations remain. For instance, the underlying causes of the observed trends—such as whether the increase in 2020 was due to direct COVID-19 complications, indirect healthcare system failures, or other factors—could not be conclusively determined. Similarly, the dramatic decline in 2022 warrants further investigation to confirm whether it signifies real progress in addressing diabetes-related mortality or is an artifact of incomplete or inconsistent data reporting. Additionally, the analysis does not account for potential regional disparities or other demographic factors that may influence mortality rates.

Despite these limitations, the findings underscore the significant impact of the COVID-19 pandemic on diabetic populations, highlighting the need for targeted public health interventions to protect vulnerable groups during global health crises. Future research with more comprehensive and granular data is recommended to address the remaining uncertainties and to develop more effective strategies for mitigating diabetes-related risks in similar scenarios. This study provides a foundation for understanding the broader implications of pandemics on chronic disease management and mortality.