Analytical Review of Healthcare Stroke Incidence & Risk Factors

Prepared by: Adaobi Cynthia Okonkwo

Tools Used: Excel, Data Visualization Techniques

Category: Healthcare Data Analytics

1. **Outline**

   This report follows a structured roadmap to present the analysis clearly and logically:

   - Introduction
   - Story of Data
   - Data Splitting and Preprocessing
   - Pre-Analysis
   - In-Analysis
   - Post-Analysis and Insights
   - Data Visualizations & Charts
   - Recommendations and Observations
   - Conclusion
   - References & Appendices

2. **Introduction**

Stroke remains a leading cause of long-term disability and mortality worldwide. Identifying the risk factors that contribute to stroke is critical for prevention, early intervention, and effective healthcare management.

This report analyzes a structured healthcare dataset containing demographic, lifestyle, medical history, and biometric variables to uncover patterns associated with stroke incidence. The workflow encompasses data cleaning, preprocessing, feature engineering, exploratory data analysis (EDA), and visualization to provide actionable insights.

The objective is to determine which factors such as age, hypertension, heart disease, BMI, glucose levels, and smoking status most strongly predict stroke risk, thereby supporting data-driven decisions in healthcare and public health initiatives.

3. **Story of Data**

- **Data Source:** The dataset is publicly available and compiled from health surveys and clinical records from Kaggle.
- **Data Collection:** Data was collected through structured medical assessments and surveys, and aggregated into a tabular format suitable for analysis in Microsoft Excel.
- **Data Structure:** Rows: Individual patient records
- **Columns:** Variables representing demographics, health status, lifestyle, and clinical measurements

- **Important Features and Their Significance:**
  - Age: Capture demographic risk trends
  - Hypertension & Heart Disease: Known clinical risk factors
  - BMI: Assess obesity-related risk
  - Avg_Glucose_Level: Indicate metabolic health
  - Smoking_status: Lifestyle-related risk indicator

Additional engineered features such as Age Group and BMI Category were added to enrich pattern

- **Limitations**:
  - Missing values in BMI
  - Potential sample bias due to survey collection
  - Limited geographic coverage

4. **Data Splitting and Preprocessing**

This stage ensured the dataset was clean, accurate, and structured for meaningful analysis.

- **Data Cleaning**

The dataset was reviewed for quality issues, including duplicates, inconsistencies, and incorrect entries. Duplicates were not present, invalid values were not present, and categorical fields were checked to improve reliability.

- **Handling Missing Values**

Missing values were addressed using appropriate strategies depending on the variable type. Numerical gaps (e.g. BMI) were filled with N/A, while missing categorical values were categorized as "Unknown" to retain record completeness.

- **Data Transformations**

Several transformations were applied to enhance analysis:

- Categorical Groupings: Age_Group and BMI_Category were created to support pattern detection.
- Risk Labels: A risk classification column was engineered using medical thresholds for glucose level, BMI, and existing conditions.

- **Data Splitting**

Variables were separated into:

- Independent Variables: Demographic, lifestyle, and biometric features such as age, hypertension, work_type, glucose level, etc.
- Dependent Variable: Stroke, representing whether the individual experienced a stroke (1 = Yes, 0 = No).

This split allowed proper correlation analysis and identification of key predictors.

- **Industry Context**

The dataset comes from the healthcare industry, focusing specifically on stroke risk analysis. Understanding the behavioral, medical, and demographic contributors to stroke helps improve preventative care.

- **Stakeholders**

Key stakeholders who benefit from these insights include: Clinical practitioners, Public health researchers, Healthcare administrators, Policy makers, Health insurance analysts.

- **Value to the Industry**

The analysis provides evidence-based insights that can:

- Support early detection of high-risk individuals
- Guide resource allocation and preventive health campaigns
- Improve patient education strategies
- Strengthen healthcare decision-making
- Reduce stroke-related mortality and long-term care costs

5. Pre-Analysis

The pre-analysis phase provides an initial understanding of the dataset and uncovers early trends before deeper statistical evaluation.

- **Identify Key Trends**

Several early patterns were observed after exploring the demographic, lifestyle, and medical variables:

- Age Distribution: A significant portion of individuals fall within the middle-aged and elderly categories, where stroke risk is naturally higher.
- Health Conditions: Hypertension and heart disease appear more common among older individuals, suggesting age-related health progression.
- Glucose Levels: Elevated average glucose levels are frequently recorded among individuals with existing chronic conditions.
- BMI Trends: Overweight and obese categories constitute a large share of the dataset, hinting at lifestyle-linked health risks.

- **Potential Correlations**

Preliminary observations suggest potential relationships worth examining in deeper analysis:

- Age vs. Stroke: Older individuals show a higher proportion of positive stroke records, indicating a probable correlation.
- Hypertension & Heart Disease vs. Stroke: Early checks show that those with either condition have a higher stroke occurrence compared to those without.
- BMI vs. Chronic Conditions: Higher BMI values coincide with increased hypertension and heart disease cases.
- Smoking Status vs. Stroke: non-smokers may show elevated risk compared to Former or current smokers.

- **Initial Insights**

Before conducting deeper modeling and statistical tests, the following insights stand out:

- Medical history is a strong early indicator of stroke likelihood (especially hypertension and heart disease).
- Age remains a critical risk factor, with stroke cases increasing sharply among older groups.
- Lifestyle factors such as smoking and BMI may influence health outcomes, suggesting they should be included as predictive variables.

- Some variables (e.g., work type, residence type) may require deeper exploration to determine their practical relevance.
- Data distribution is imbalanced, stroke cases may be significantly fewer than non-stroke cases, which is typical in medical datasets. This will affect interpretation and may require technique adjustments.

6. In-Analysis

This stage represents the core of the analytical process, where deeper exploration of the dataset reveals verified and emerging patterns. Using Excel-based techniques such as PivotTables, correlation checks, conditional formatting, and lookup functions, several insights were uncovered.

- **Unconfirmed Insights (Hypotheses Under Evaluation)**

During initial exploration, several patterns appeared promising but require further validation with statistical tools:

- Chronic conditions and stroke: Early pivot results suggest individuals with hypertension or heart disease appear more frequently among stroke cases. However, this needs further correlation analysis to confirm strength and direction.
- High glucose levels: Individuals within the "High" or "Critically High" glucose categories seem to show elevated stroke incidence. This hypothesis needs confirmation using mean comparison and segmentation.
- Smoking status: Preliminary checks show that *formerly smoking* individuals might display higher stroke occurrence than current smokers or non-smokers, though this may be influenced by age distribution.
- BMI and stroke risk: Higher BMI categories (Overweight/Obese) appear more often in chronic-condition flags, but direct links to stroke are not yet statistically validated.

- **Recommendations (Preliminary and Based on Emerging Patterns)**

Although the analysis is ongoing, early patterns allow for initial recommendations:

- Prioritize screening for individuals with hypertension or heart disease, as these groups show consistently higher stroke occurrence.

- Introduce targeted monitoring for patients with elevated glucose levels, since high glucose may be a strong early warning indicator.
- Encourage lifestyle interventions for overweight and previously-smoking individuals, as they show evidence of elevated health risks.
- Implement risk stratification models within healthcare settings using core variables like Age, Hypertension, Glucose, and BMI to support prevention programs.

These recommendations will be refined once hypotheses are fully validated.

- **Analysis Techniques Used in Excel**

Excel served as the primary tool for exploring and validating relationships in the dataset. Key functionalities included:

- Pivot Tables: Used to segment stroke cases by demographic and medical variables (Age Groups, BMI Categories, Glucose Levels) for pattern detection.
- Correlation Checks: Utilized Excel's correlation functions and data analysis add-in to identify relationships between numeric variables such as Age, BMI, and Glucose Level.
- Pivot Charts: Visual forms of pivot outputs helped compare variable distribution and stroke presence across categories.

7. Post-Analysis and Insights

This stage consolidates the findings from the full analytical process, confirming or rejecting the initial hypotheses and interpreting the patterns observed within the dataset. The validated insights offer a deeper understanding of the factors associated with stroke incidence and provide an evidence-based foundation for recommendations.

- **Key Findings**

Based on the completed analysis, several significant insights emerged:

- Age is the strongest predictor of stroke incidence. Stroke cases increase sharply among individuals aged 50 and above, its highest peak between ages 70 to 80, confirming age as a high-risk factor.

- Hypertension and heart disease show strong positive associations with stroke. Individuals with either condition and especially those with both appear disproportionately in the stroke group.
- High glucose levels correlate with elevated stroke occurrence. Average glucose levels above 140 mg/dL were more common among stroke patients.
- BMI alone shows a weaker relationship, although individuals categorized as Obese had a slightly higher representation in stroke cases compared to Normal and Overweight categories.
- Smoking status reveals an unexpected trend: Never Smoked individuals displayed higher stroke incidence than current smokers or those who formerly smoked.
- Work type and marital status show minor but observable trends, with individuals in private-sector and married categories slightly overrepresented in stroke cases.

- **Comparison with Initial Findings**

Comparing the validated results with the early hypotheses reveals several aligned and surprising outcomes:

- Confirmed Expectations: The initial assumption that age, hypertension, and heart disease are major stroke predictors was fully supported. Hypothesis that high glucose levels contribute to stroke risk was validated through comparative pivot summaries and segmentation.
- Partially Confirmed Assumptions: BMI was expected to show a stronger relationship with stroke risk; however, the analysis revealed only a mild association.
- Unexpected or Counter-Intuitive Results: The higher stroke incidence among never smoked individuals contradicted the expectation that current smokers would be most at risk.
- Neutral or Low-Impact Findings: Work type and residence type had relatively weak influence and did not significantly impact stroke distribution.

8. Data Visualizations & Charts

Visual representations were created in Microsoft Excel to simplify complex data relationships and highlight key patterns related to stroke risk. These charts provide quick, intuitive insights for stakeholders and support data-driven decision-making. A summary of the visualizations and their interpretations is presented below.

- **Visuals**
  - **Age Distribution vs. Stroke Outcomes (Histogram Chart)**

A histogram chart was used to compare age groups with stroke incidence.

Insight:

The visualization shows a sharp increase in stroke cases in the 70 – 80 category, confirming age as one of the most influential risk factors. Younger age groups show minimal stroke representation.

- **Age Group Distribution of Hypertension and Heart Disease (Clustered Bar Chart)**

This chart visualizes the distribution of hypertension and heart disease among Age Group.

Insight:

The chart clearly shows the prevalence of hypertension and heart disease increases significantly with age. Among older individuals, there are 339 cases of hypertension and 232 cases of heart disease, making this group the most affected.

- **Stroke Prevalence across Glucose_Level_Category (Pie Chart)**

This visualization categorizes average glucose levels (e.g., Normal, Prediabetic, Diabetic) and compares them with stroke cases.

Insight:

A majority of stroke cases occurred among individuals with normal glucose levels (156 cases), which may indicate that glucose level alone is not a sufficient predictor of stroke risk.

- **BMI Category Distribution among Stroke and Non-Stroke Patients (Clustered Column Chart)**

BMI categories (Underweight, Normal, Overweight, Obese) were plotted against stroke and non-stroke outcomes.

Insight:

The majority of stroke cases are observed among individuals classified as Obese (98 cases) and Overweight (75 cases), together accounting for over 85% of all strokes in the dataset. However, it's also worth noting that obesity has the largest population count (1,920 individuals), which may partly explain the higher stroke numbers.

- **Stroke Prevalence across Smoke_Status (Column Chart)**

This visualization shows the proportions of Never Smoked, Formerly Smoked, and Smokes categories across stroke outcomes.

Insight:

The highest number of stroke cases occurred among individuals who never smoked (90 cases), followed by those who formerly smoked (70 cases). Interestingly, those with unknown smoking status (47 cases) accounted for almost as many strokes as active smokers (42 cases). This suggests that while smoking is a known risk factor, stroke incidence is influenced by multiple factors beyond smoking alone, and data gaps (unknown status) may also hide important patterns.

- **Proportion of Stroke and Non-Stroke Cases by Gender**

This visualization shows the proportions of Gender by Stroke and Non-Stroke cases.

Insight:

The dataset shows that among individuals who did not experience a stroke, females (2,853) slightly outnumber males (2,007). For those who experienced a stroke, females (141) also recorded a higher number of cases compared to males (108). No stroke cases were reported under the "Other" gender category. This suggests that while females make up a larger proportion of the population overall, they also show a slightly higher absolute count of stroke cases than males.

- **Impact of Work Type on Stroke Cases (Column Chart)**

This visualization shows the proportions of Private, Self-employed, Govt_job, Children, Never_worked categories across stroke outcomes.

Insight:

Most stroke cases are concentrated among individuals in the Private sector (149 cases) and the Self-employed group (65 cases), together accounting for the majority of all strokes. Govt_job workers reported fewer cases (33), while children (2) and Never_worked (0) had negligible counts.

- **Impact of Marriage on Stroke (Column Chart)**

This visualization shows the effect of Marriage categories across stroke outcomes.

Insight:

The majority of stroke cases (220 out of 249) are among individuals who are married, while only 29 cases are from those who have never married. This suggests that strokes are more prevalent in older or married populations, possibly due to age-related health risks.

- **Stroke Cases by Residence Type (Column Chart)**

This visualization shows the proportions of Urban and Rural categories across stroke outcomes.

Insight:

Stroke cases are slightly higher among individuals living in urban areas (135 cases) compared to those in rural areas (114 cases). This suggests that strokes affect people in both settings fairly equally, though urban residents appear to be at a marginally greater risk.

- **Health Care's Stroke Incidence and Risk Factors Analysis (Excel Dashboard)**

A consolidated dashboard was developed using slicers, pivot charts, and KPI indicators to allow interactive exploration of: Age Distribution vs. Stroke Outcomes, Age Group Distribution of Hypertension and Heart Disease, Stroke Prevalence across Glucose_Level_Category, BMI Category Distribution among Stroke and Non-Stroke Patients, Stroke Prevalence across Smoke_Status, Proportion of Stroke and Non-Stroke Cases by Gender**.**

Dashboard Insights:

The dashboard provides an at-a-glance view of how demographic and biometric variables interact with stroke presence, helping stakeholders quickly identify high-risk populations.

9. Recommendations and Observations

This section translates the analytical findings into practical, data-driven recommendations intended to support healthcare practitioners, policy makers, and hospital administrators in improving stroke prevention and management strategies.

- **Actionable Insights**
  - Prioritize High-Risk Age Groups: Data shows that individuals aged **50 years and above** have the highest likelihood of experiencing a stroke.

Recommendation: Implement targeted health monitoring programs for elderly patients, including regular neurological checks, glucose testing, and blood pressure screening.

- Strengthen Hypertension & Heart Disease Management Programs: Chronic conditions (hypertension and heart disease) appear as the strongest predictors of stroke. **Recommendation:** Develop or expand chronic disease management clinics to provide continuous counseling, medication adherence reminders, and lifestyle intervention support.
- Enhance Public Awareness Campaigns on Lifestyle Risks: Although, Smoking status does not contribute significantly to stroke risk.
  Recommendation: Launch community health education campaigns focused on the long-term effects of smoking, alcohol consumption, sedentary lifestyles, and poor diet.

- **Optimizations or Business Decisions**

Allocate More Resources to Preventive Care Units: Given the strong correlation between chronic conditions and stroke, shifting more funding and staffing towards preventive care could reduce long-term patient burden and hospital costs.

Deploy Digital Health Tools for Monitoring High-Risk Patients: Using patient dashboards, automated reminders, and telehealth check-ins can reduce the likelihood of medical emergencies among vulnerable groups.

Enhance Patient Segmentation for Efficient Care Delivery: Segregating patients by risk categories (e.g., Age Group, BMI Category) can help healthcare facilities personalize treatment and optimize resource allocation.

Establish a Stroke Early Warning System (EWS): Integrating age, BMI level, and chronic-condition data into an early warning system could support early detection and timely interventions.

- **Unexpected Outcomes & Observations**

- Non-Smokers (Never Smoked) Showing Higher Stroke Rates Than Current Smokers:

This trend appears counterintuitive because smoking is a well-established medical risk factor for stroke. Several factors may explain this pattern in the dataset:

Underreporting of smoking habits: Some individuals categorized as Never Smoked may have inaccurately reported their smoking history, leading to misclassification.

Health-driven behavior change: Current smokers may be younger or may not yet have accumulated enough long-term exposure for stroke risk to become visible.

Dataset imbalance: If the "Current Smoker" category has fewer respondents, even a small number of strokes in the "Never Smoked" group can skew comparisons.

Uncaptured lifestyle factors: Other risks, such as diet, stress, physical inactivity, may be more pronounced in the non-smoker subgroup.

- **Glucose Level Category Showing Weaker Predictive Power**

Despite medical literature linking elevated glucose levels (particularly in diabetic patients) to higher stroke risk, the dataset shows weaker predictive strength for this variable. Potential explanations include:

Wide variation within categories: Categories such as "Normal" or "High" may contain patients with overlapping glucose ranges, reducing clear predictive separation.

Single-time measurement limitation: A one-time glucose reading does not capture chronic patterns; long-term glucose indicators (e.g., HbA1c) were not included.

Dataset bias: If stroke cases are not evenly distributed across glucose categories, predictive strength becomes limited.

- **Some Individuals Developing Stroke Despite No Chronic Conditions**

This suggests that genetic predispositions, stress, environmental factors, or undiagnosed conditions may play roles not captured in the dataset.

10. **Conclusion**

This analytical project provided a comprehensive assessment of stroke incidence and the influence of key demographic, lifestyle, and clinical variables within the dataset. Through a structured workflow involving data cleaning, preprocessing, feature engineering, visualization, and interpretation, the analysis revealed clear patterns especially the strong predictive weight of age, hypertension, heart disease, smoking status, and glucose levels. These insights underscore the importance of early screening and targeted intervention strategies in healthcare environments.

- **Key Learnings**

The analysis confirmed that stroke risk increases substantially with age, particularly among individuals above 50 years. Cardiovascular-related conditions such as hypertension and heart disease emerged as the strongest contributors to stroke likelihood within the dataset. Lifestyle factors, including smoking status and BMI category did not demonstrate significant associations. Some results, such as higher stroke rates in never-smokers, revealed potential data inconsistencies or underlying demographic effects. Feature engineering, such as creating Age Groups, BMI Categories, and Glucose Level Category, further improved interpretability and enhanced the visibility of high-risk subpopulations.

- **Limitations**

Several constraints may influence the reliability and generalizability of the findings.

- Data completeness issues: The dataset contained missing values for variables such as BMI and smoking status, potentially affecting category accuracy.
- Sampling bias: Some subgroups (e.g.,certain work types) were underrepresented, contributing to unexpected statistical patterns.
- Lack of temporal data: Without multi-year health records, trends over time, disease progression, and long-term risk factors could not be examined.
- No medical validation: Insights are derived from statistical patterns, not clinical diagnoses, which restricts direct healthcare decision-making.

- **Future Research**

To strengthen predictive performance and expand healthcare applicability, several avenues for further research are recommended:

- Integration of longitudinal medical data (e.g., multi-year glucose records, blood pressure trends, and medical histories).
- Inclusion of clinical biomarkers such as cholesterol levels, family health history, HbA1c, diet, and physical activity metrics.
- Stratified subgroup analysis to better understand unexpected outcomes (e.g., non-smoker stroke rates).

- Expansion of demographic diversity to reduce bias and improve real-world generalization across populations.

Overall, this project demonstrates how structured data analysis can uncover meaningful healthcare insights and support early detection initiatives. With more complete datasets and expanded variables, future explorations can yield even stronger predictive models and more precise recommendations for stroke prevention and risk management.