Stroke Risk Analysis

Data-Driven Analysis of Stroke Risk Factors

Prepared For: Prospective Employers - Healthcare & Analytics Sector

Author: John Paul Medina

Date: April 20, 2025

Github: https://www.github.com/jmedinacs

LinkedIn: Https://www.linkedin.com/in/jpmedinacs

1.0 Executive Summary

This report analyzes the leading predictors of stroke using patient health and demographic data. Through exploratory analysis, statistical testing, and selected multivariate comparisons, it identifies **age, heart disease, hypertension, and glucose level** as the most significant individual risk factors. When combined, these factors greatly amplify stroke risk with specific subgroups, such as **elderly patients with both heart disease and diabetes**, reaching a stroke likelihood of **40%**, which nearly eight times the population stroke risk baseline of **4.87%**.

Based on these findings, this report recommends **early patient education** about compounding risk factors, **enhanced screenings** starting in middle age, **targeted outreach** to high-risk subgroups, and the **use of machine learning** to scale risk assessment and personalize patient healthcare plans. These insights aim to support data-driven decisions in healthcare screening and preventative care initiatives.

1.1 Key Findings

- When taken as individual risk factors, the top four features that increases a patient's risk of experiencing a stroke are:
    - Age (elderly) - 21.51%
    - Heart Disease - 17.03%
    - Hypertension - 13.25%
    - Glucose Level (diabetes) - 10.00%

  These values are approximately **two to four times higher** than the baseline stroke risk of **4.87%** of the total population of the dataset.
- Among the features that are included in this analysis, the **Chi-squared test** revealed that age, heart disease, hypertension, diabetes, marriage

status, work type, bmi, and smoking status are all **statistically significant features** that contribute to the risk of a patient experiencing a stroke.

- The **bivariate analysis** (pairwise combination of features) revealed that:
    - **Elderly patients with heart disease** had the **highest** observed stroke risk of 27.27%.
    - **Elderly with hypertension** is at 26% risk.
    - **Heart disease and diabetes** at 25% risk.
    - **Elderly with diabetes** at 21.67% risk.
    - **Heart disease and hypertension** at 20.31% risk.
    - **Hypertension with diabetes** at 17.02% risk.
- A coursory exploration of **multivariate analysis** (three or more features) revealed that **elderly patients with both heart disease and diabetes face a 40% risk** of experiencing a stroke, which is the highest observed risk in this project. Due to the **exponential growth** in subgroup combinations, a complete multivariate analysis involving all features and categories would require **machine learning or statistical modeling** to obtain a more comprehensive and scalable understanding of how three or more compounding risk factors affect the patient's stroke risk.

1.2 Recommendations

- **Initiate patient education on stroke risk factors starting in young adulthood** Promote early awareness and preventative habits before high-risk conditions develop.
- **Enhance patient screening and offer dietary and lifestyle support beginning in middle age**

  Focus especially on individuals with risk factors such as hypertension, heart disease, or diabetes.

- **Target high-risk subgroups — particularly elderly patients with multiple risk factors**
  Prioritize outreach, screenings, and personalized care for patients exhibiting compounded risk (e.g., elderly with both heart disease and diabetes, who face up to a 40% stroke risk).
- **Leverage machine learning or statistical modeling**
  Use predictive tools to assess stroke risk from multiple interacting features, allowing healthcare providers to deliver more precise and data-driven care.

2.0 Introduction

This analysis examines how various risk factors including age, heart disease, hypertension, glucose level, and smoking influence the likelihood of a patient experiencing a stroke. The report identifies the top predictors associated with increased stroke risk, explores how combinations of two risk factors compound that risk, and discusses the potential for using machine learning to conduct more robust multivariate analysis of feature combinations beyond two. These insights aim to support data-driven recommendations for patient screening and preventive healthcare planning.

2.1 Objectives

- **Identify** the top predictors associated with increased risk of experiencing a stroke.
- **Explore** how combinations of two risk factors compound stroke risk.
- **Discuss** the potential of using machine learning and predictive modeling to deepen understanding and enable healthcare providers to deliver more personalized, data-driven care.

2.2 Data Overview

This analysis uses the **Stroke Prediction Dataset** published by fedesoriano on Kaggle, which contains anonymized health data for **5,110 patients**. The dataset includes a mix of demographic, lifestyle, and medical attributes that may influence stroke risk.

Key features include:

- **Demographics**: age, gender, marital status, residence type, work type
- **Health indicators**: hypertension, heart disease, average glucose level, BMI, smoking status
- **Target variable**: whether the patient has previously experienced a stroke

Several features were standardized or transformed to improve analysis:

- **Age**, **glucose level**, and **BMI** were binned into medically relevant categories
- Binary features like **hypertension** and **heart disease** were recoded for clarity
- Categorical fields were normalized to lowercase for consistency

The dataset was cleaned and prepared for analysis in Google Sheets, with all data transformations and decisions logged in the cleaning log.

2.3 Preprocessing Summary

The dataset underwent several preprocessing steps to prepare it for analysis. These steps ensured consistency, interpretability, and improved support for categorical comparisons and visualization:

- **Missing values**: All empty cells were checked; only the BMI column contained missing values, which were imputed using the **median BMI** (28.1) to minimize outlier influence.
- **Categorical standardization**: All categorical fields (e.g., gender, work type, residence type) were converted to lowercase and formatted consistently. Verbal categories were used for binary variables (e.g., “have hypertension” instead of 1).
- **Column recoding and labeling**:
    - Binary indicators for stroke, heart disease, and hypertension were recoded into descriptive labels.
    - One "other" entry in gender was retained for integrity but excluded from gender-based analysis.
- **Feature binning**:
    - **Age** was grouped into five medically-relevant categories: pediatric, young adult, middle-aged adult, senior, and elderly.
    - **BMI** was binned into underweight, normal, overweight, and obese.
    - **Glucose level** was binned into hypoglycemic, normal, prediabetic, and diabetic, assuming fasting measures based on CDC, ADA, and WHO guidelines.
- **Helper columns** were added to support sorting and display of ordered categorical data (e.g., numeric order for age or BMI groups).
- **Outlier handling**: High values in glucose and BMI were identified using the IQR method but retained, as they reflected plausible medical conditions.

All cleaning steps were documented in a structured log, and decisions were tracked to support transparency and reproducibility. To view the complete log for the project, follow this log link.

3.0 Exploratory Data Analysis (EDA)

This section documents the analysis process and the key insights gained. The EDA proceeds through the following stages:

- **Feature vs. Stroke Analysis** – Examines how each individual feature and its categories affect the likelihood of stroke. This includes both overall feature distribution and comparisons between feature categories and stroke occurrence.
- **Chi-Squared Test** – Conducts chi-squared tests of independence to determine which features are statistically significant in contributing to stroke occurrence. Additional analysis identifies which categories within these features contribute most strongly.
- **Bivariate vs. Stroke Analysis** – Performs pairwise (two-feature) analysis to evaluate how combinations of two risk factors impact stroke risk compared to individual effects.
- **Multivariate Samples** – Presents two multivariate examples (three features vs. stroke) to illustrate how compounding risk factors behave, and to demonstrate the value of a machine learning–based approach for more comprehensive predictive modeling.

3.1 Feature vs Stroke Analysis

Each feature in the dataset was analyzed for both its overall distribution and its relationship to stroke occurrence. **Figure 1** illustrates the analytical process applied to each feature. The first chart evaluates the **distribution of patient records by category** (e.g., age group), assessing the sample size, shape, and skew to determine the strength and reliability of each category. The second chart displays the **percentage of patients with and without stroke** across those same categories.

After completing the feature-by-feature analysis, a **top risk predictor summary** was created to identify which categories most significantly increase the risk of stroke. **Figure 2** shows a sorted and visualized summary of the stroke risk associated with each high-risk category. The top five are as follows:

- **Elderly (age)** – 21.51%
- **Heart disease** – 17.03%
- **Hypertension** – 13.25%
- **Senior (age)** – 11.85%
- **Diabetic (glucose level)** – 10.00%

These values are approximately **two to four times higher** than the **baseline stroke risk of 4.87%**, calculated from the overall proportion of stroke cases in the dataset.

Figure 1: Age Distribution and Age vs. Stroke Rate

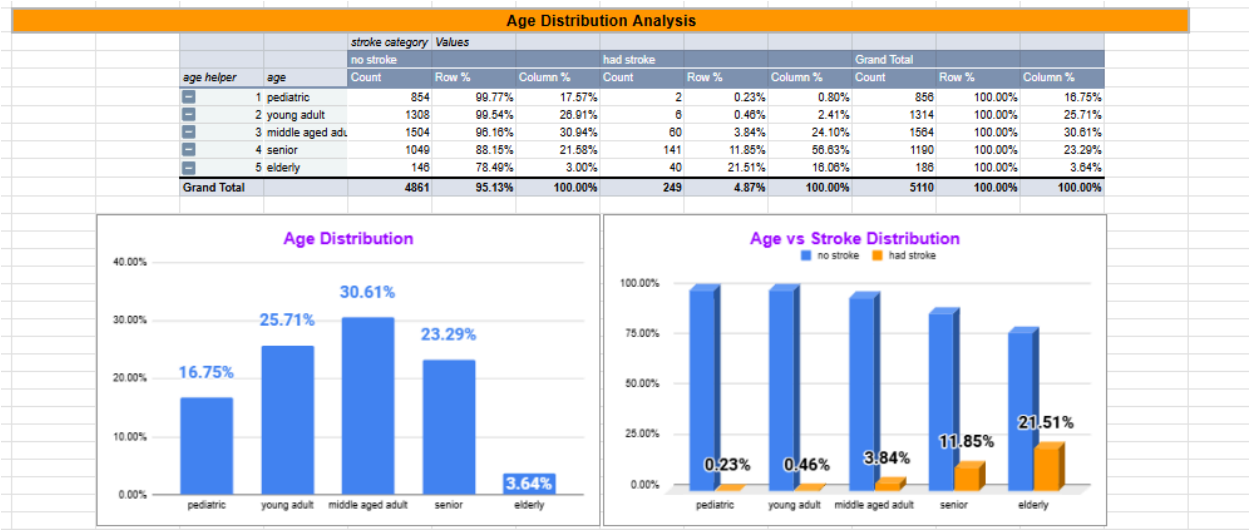
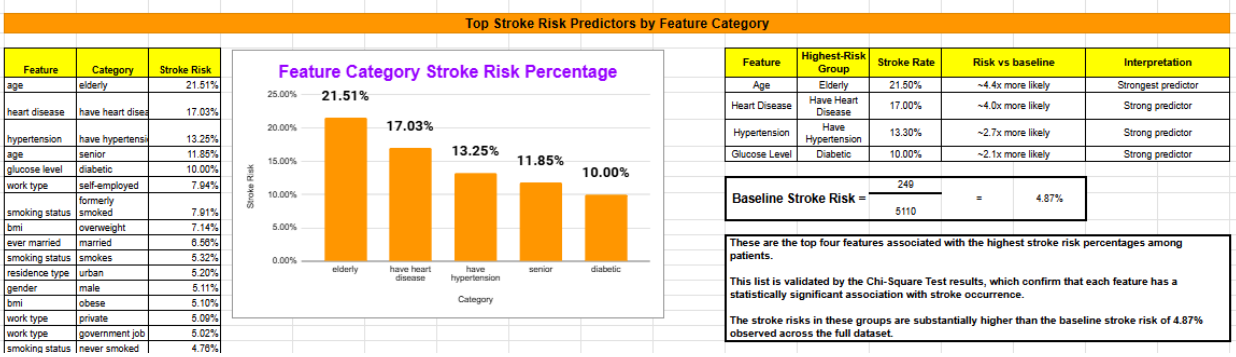


Figure 2: Top Stroke Risk Predictors Summary (Feature vs. Target Analysis)



Top Stroke Risk Predictors by Feature Category

Feature	Category	Stroke Risk
age	elderly	21.51%
heart disease	have heart disea	17.03%
hypertension	have hypertensi	13.25%
age	senior	11.85%
glucose level	diabetic	10.00%
work type	self-employed	7.94%
smoking status	formerly smoked	7.91%
bmi	overweight	7.14%
ever married	married	6.56%
smoking status	smokes	5.32%
residence type	urban	5.20%
gender	male	5.11%
bmi	obese	5.10%
work type	private	5.09%
work type	government job	5.02%
smoking status	never smoked	4.76%

Feature	Highest-Risk Group	Stroke Rate	Risk vs baseline	Interpretation
Age	Elderly	21.50%	~4.4x more likely	Strongest predictor
Heart Disease	Have Heart Disease	17.00%	~4.0x more likely	Strong predictor
Hypertension	Have Hypertension	13.30%	~2.7x more likely	Strong predictor
Glucose Level	Diabetic	10.00%	~2.1x more likely	Strong predictor

Baseline Stroke Risk = 249 / 5110 = 4.87%

These are the top four features associated with the highest stroke risk percentages among patients.

This list is validated by the Chi-Square Test results, which confirm that each feature has a statistically significant association with stroke occurrence.

The stroke risks in these groups are substantially higher than the baseline stroke risk of 4.87% observed across the full dataset.

3.2 Chi-Squared Test

Since the continuous features in the dataset (e.g., age, glucose level, BMI) were converted into medically standardized categorical bins — and all other relevant features were already categorical — the **Chi-Squared Test of Independence** was used to test the association between each feature and stroke occurrence.

This statistical test evaluates whether there is a **significant relationship between two categorical variables** by comparing the observed frequency distribution to what would be expected under independence. In this case, it measures whether stroke occurrence is independent of each feature's categories.

Figure 3 illustrates the formulas used in the chi-squared test and provides a sample calculation applied to the **age** feature. For each feature, expected values, chi-squared contributions, and p-values were computed. A p-value below 0.05 indicates a statistically significant relationship with stroke. Contributions above **100** or representing more than **50% of the total statistic** were flagged as strong indicators and analyzed further.

Figure 4 presents the summary results of the chi-squared tests across all features, including p-values and the most influential categories. The results

confirm that **age, heart disease, hypertension, and glucose level** are the strongest individual predictors of stroke, each contributing significantly to the overall chi-square statistic.

By contrast, categories such as **“never married”** and **“children”** showed **protective associations**, with stroke rates significantly lower than expected; likely influenced by age-related confounding. Features such as **BMI** and **smoking status** were moderate contributors, while **gender** and **residence type** were not statistically significant.

The accompanying chart in **Figure 4** visualizes the **direction and strength of contribution** for each category, supporting prioritization of risk factors in future predictive modeling and public health outreach.

Figure 3: Formulas and Age vs Stroke Chi-Squared Test

FORMULAS						
Expected Count =	Row Total	x	Column Total / Grand Total			
Chi-Square =	(Observed	-	Expected) / Expected			
Degrees of Freedom =	(# of rows - 1)	*	(# of columns -1)			
p - value =	CHISQ.DIST.RT (chi_sqr_stat, degrees of freedom)		p - value < 0.05 is significant			

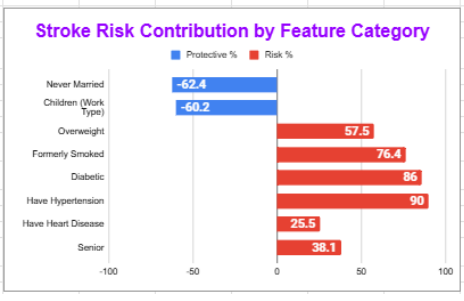
Age vs Stroke Chi-Squared Test							
Age Category	no stroke		had stroke		Row Total	Chi Contribution	
	observed	expected	observed	expected		no stroke	had stroke
elderly	146	176.9365949	40	9.063405088	186	5.409129215	105.5974985
middle aged adult	1504	1487.789432	60	76.21056751	1564	0.1766261363	3.448111039
pediatric	854	814.2888454	2	41.7111546	856	1.9366295	37.80705221
senior	1049	1132.013699	141	57.98630137	1190	6.087624353	118.8431405
young adult	1308	1249.971429	6	64.02857143	1314	2.693913657	52.59082042
Column Total	4861		249		5110		
DF	4						
chi-square stat	334.5905455		SIGNIFICANT				
p-val	0.000000000						

Age had the highest total chi-square value, with strong contributions from senior and elderly patients, and a notably lower-than-expected stroke rate among young adults.

While no single age group contributed more than 40% of the total statistic, the combined variation across age groups makes age the most robust predictor of stroke risk in the dataset.

Figure 4: Chi-Squared Test Summary Table and Visualization

Chi-Square Test Summary Table						
Feature	Chi-Square Stat	p-value	Top Contributing Group	Top Contribution	% of Total χ^2	Interpretation
Age	334.59	< 0.000001	Senior (Stroke)	118.84	35.50%	Strongest Predictor
Heart Disease	93.01	< 0.000001	Heart Disease (Stroke)	83.7	90.00%	Strong Predictor
Hypertension	83.6	< 0.000001	Hypertension (Stroke)	71.77	86.00%	Strong Predictor
Glucose Level	70.58	< 0.000001	Diabetes (Stroke)	53.95	76.40%	Strong Predictor
Ever Married	59.98	< 0.000001	Never Married (Protective)	37.44	62.40%	Protective – significantly fewer strokes than expected
Work Type	49.16	< 0.000001	Children (Protective)	29.6	60.20%	Not a standalone predictor (age confounded)
BMI	44.69	< 0.000001	Overweight (Stroke)	17.03	38.10%	Moderate Contributor
Smoking Status	29.15	< 0.000002	Former Smoker (Stroke)	16.75	57.50%	Moderate Contributor
Gender	0.42	-	-	-	-	Not Significant
Residence Type	1.22	-	-	-	-	Not Significant



3.3. Bivariate versus Stroke Analysis

It is not uncommon for multiple risk factors to be present in a patient's medical history. Understanding how combinations of risk factors compound stroke risk is essential for targeted prevention.

As an illustrative example, Figure 5 explores the combined effect of **heart disease and diabetes**. Patients with both conditions experienced a **25% stroke rate**, compared to just **4.18%** for patients with neither. This analysis highlights how overlapping risk factors can lead to dramatically higher stroke likelihood.

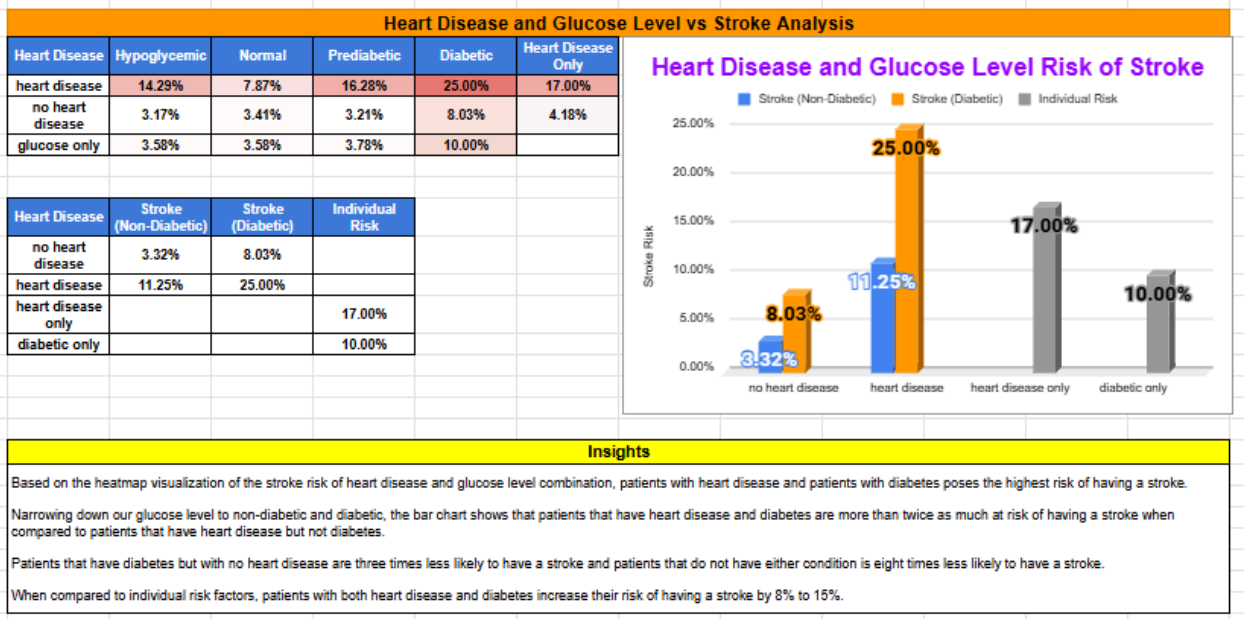
While numerous other combinations were explored during the analysis phase, only a **representative subset** is presented here to avoid overwhelming the reader. These examples were selected based on the strength of their insights and relevance to real-world prevention strategies.

Figure 6 shows the stroke risk of pairing the top four risk factors, and the result is as follows:

- **Elderly with heart disease** - 27.27%
- **Elderly with hypertension** - 26%
- **Heart disease and diabetes** - 25%
- **Elderly and diabetes** - 21.67%
- **Heart disease and hypertension** - 20.31%
- **Hypertension and diabetes** - 17.02%

For the complete bivariate visualization and analysis, view the interactive sheet here: Stroke Risk Bivariate Analysis Sheet

Figure 5: Heart Disease and Diabetes vs Stroke Analysis



Heart Disease and Glucose Level vs Stroke Analysis

Heart Disease	Hypoglycemic	Normal	Prediabetic	Diabetic	Heart Disease Only
heart disease	14.29%	7.87%	16.28%	25.00%	17.00%
no heart disease	3.17%	3.41%	3.21%	8.03%	4.18%
glucose only	3.58%	3.58%	3.78%	10.00%	

Heart Disease	Stroke (Non-Diabetic)	Stroke (Diabetic)	Individual Risk
no heart disease	3.32%	8.03%	
heart disease	11.25%	25.00%	
heart disease only			17.00%
diabetic only			10.00%

Insights

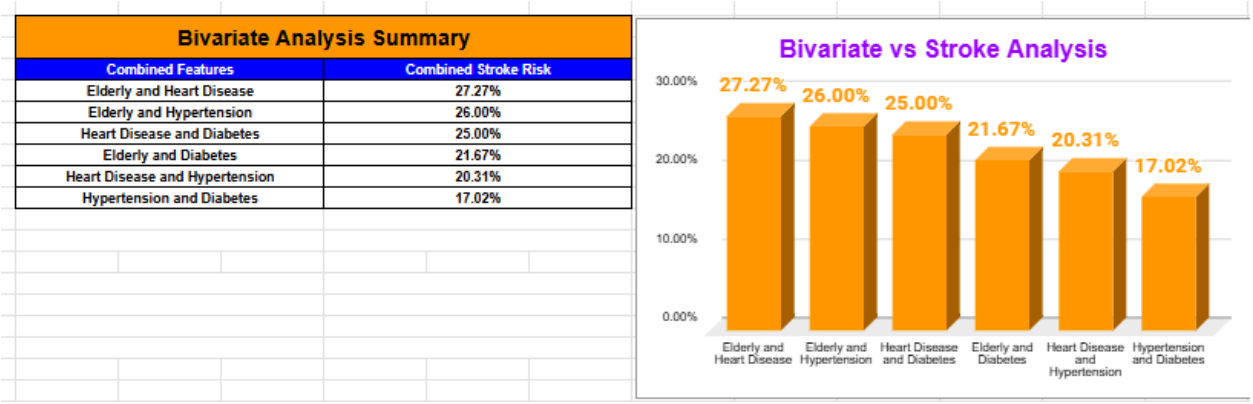
Based on the heatmap visualization of the stroke risk of heart disease and glucose level combination, patients with heart disease and patients with diabetes poses the highest risk of having a stroke.

Narrowing down our glucose level to non-diabetic and diabetic, the bar chart shows that patients that have heart disease and diabetes are more than twice as much at risk of having a stroke when compared to patients that have heart disease but not diabetes.

Patients that have diabetes but with no heart disease are three times less likely to have a stroke and patients that do not have either condition is eight times less likely to have a stroke.

When compared to individual risk factors, patients with both heart disease and diabetes increase their risk of having a stroke by 8% to 15%.

Figure 6: Bivariate Analysis Summary Table and Visualization



Bivariate Analysis Summary

Combined Features	Combined Stroke Risk
Elderly and Heart Disease	27.27%
Elderly and Hypertension	26.00%
Heart Disease and Diabetes	25.00%
Elderly and Diabetes	21.67%
Heart Disease and Hypertension	20.31%
Hypertension and Diabetes	17.02%

3.4 Multivariate Samples

Multivariate analysis involves combining three or more features to examine how they interact to affect the target variable, in this case, stroke occurrence. While this type of analysis is crucial for uncovering complex risk patterns, it typically requires more sophisticated methods than simply testing every possible feature combination.

Two sample combinations were analyzed:

1. **Elderly patients** by **heart disease** and **hypertension** status
2. **Elderly patients** by **heart disease** and **diabetes** status

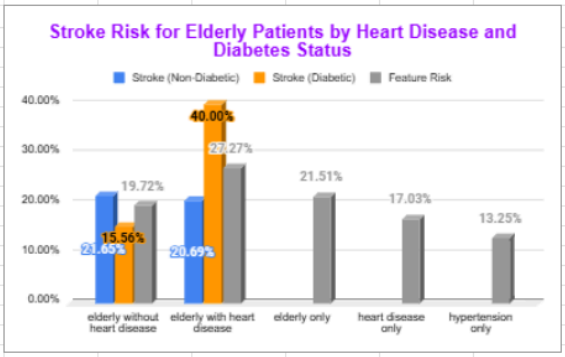
In the first combination, **elderly patients** grouped by their **heart disease** and **hypertension** status showed a highest stroke risk of **28.57%**. This analysis was limited to patients who are elderly and either had or did not have heart disease, and who were either hypertensive or not, this accounts for just **4 out of 20 possible combinations** if all categories across the three features were considered.

As shown in **Figure 8**, the second example examined **elderly patients** categorized by **heart disease** and **glucose level** (binary: diabetic or non-diabetic). The group with **all three risk factors**, elderly, heart disease, and diabetes showed a stroke risk of **40%**. Again, this subset represents only **4 of 20 possible combinations** when restricted to elderly patients and diabetic vs non-diabetic. If we expand the analysis to include **all age groups**, along with **binary heart disease** status and **multiple glucose level categories**, the number of combinations increases to **40**.

Thus, both of these samples represent only a quarter or less of their respective combination spaces. To comprehensively analyze the effects of three or more interacting risk factors on stroke occurrence, **predictive modeling techniques such as machine learning** are more appropriate and scalable than manual cross-tabulation.

Figure 8: Stroke Risk for Elderly Patients by Heart Disease and Diabetes Status

Stroke Risk for Elderly Patients by Heart Disease and Diabetes Status			
Feature Combination	**Stroke (Non-Diabetic)**	**Stroke (Diabetic)**	**Feature Risk**
elderly without heart disease	21.65%	15.56%	19.72%
elderly with heart disease	20.69%	40.00%	27.27%
elderly only			21.51%
heart disease only			17.03%
hypertension only			13.25%



Insight and Comment

Among elderly patients, those who have both heart disease and diabetes face a 40.00% risk of stroke — the highest observed stroke rate across all subgroup combinations in this analysis. This is nearly 8 times the overall population stroke rate of 4.87%, and around double the risk of the compounding risk factors in this comparison.

Interestingly, stroke risk does not increase uniformly for all combinations. For example, elderly patients without heart disease show a slightly lower stroke risk when diabetic (15.56%) compared to non-diabetic (21.65%), which may reflect smaller sample sizes or other confounding variables.

This example highlights how compounding risk factors can lead to extremely elevated stroke risk, but also reinforces the challenge of manually analyzing higher-order combinations. This 3-feature comparison covers only 4 of 20 possible combinations between age, heart disease, and a binary glucose level (diabetic vs non-diabetic).

Exploring the full set of glucose categories would expand the analysis to 40 combinations, further increasing complexity and reinforcing the need for statistical modeling or machine learning to handle the full scope of interactions.

To fully explore these relationships and control for interactions between multiple variables, statistical modeling or machine learning techniques would be required.

4.0 Recommendations

Based on the exploratory and statistical analysis of stroke risk factors, the following recommendations are proposed to guide public health campaigns, early screening efforts, and future research initiatives:

1. **Initiate Stroke Risk Education Early**

   Begin patient education on stroke risk factors in **young adults** to **build awareness** before chronic high-risk conditions emerge. Promoting healthy habits, regular exercise, and routine screenings early on can play a preventive role in **reducing long-term risks**.

2. **Enhance Midlife Screening and Support**

   Encourage primary care providers to offer more comprehensive stroke risk screenings starting in middle age, especially for patients with known conditions such as **hypertension, diabetes, or heart disease**. Integrating lifestyle counseling and support services into routine care can help manage or delay the progression of these conditions and prevent additional risk factors from compounding.

3. **Target High-Risk Subgroups for Outreach and Care**
   Prioritize outreach efforts, personalized care plans, and preventative screenings for **elderly patients with multiple risk factors**. Our analysis showed that patients who are elderly and have both **heart disease and diabetes** face up to a **40% likelihood of stroke**, more than **eight times** the baseline risk of **4.87%**. These compounded risk cases present a critical opportunity for timely intervention.
4. **Leverage Predictive Modeling for Risk Stratification**
   Manual analysis becomes insufficient when evaluating multiple interacting risk factors beyond simple pairings. Healthcare systems should consider adopting **machine learning models or logistic regression tools** to assess complex stroke risk profiles. These tools can help clinicians stratify patients by risk level and deliver more **personalized, proactive care** that tailored to their **demographic and medical history.**

5.0 Conclusion

This analysis set out to identify the leading predictors of stroke using historical patient data, with the goal of supporting more effective awareness campaigns, screening strategies, and personalized care plans. Through exploratory data analysis, chi-squared testing, and targeted multivariate exploration, several high-risk groups were identified; most notably elderly patients with chronic conditions such as heart disease, hypertension, or diabetes.

The findings underscore the importance of early prevention, midlife intervention, and targeted care for high-risk individuals. While single risk factors are associated with elevated stroke likelihood, our analysis also revealed how **compounded risks** can dramatically increase stroke probability; up to **40%** in some subgroups.

To fully leverage these insights in clinical practice, healthcare systems must continue evolving their data strategies. Predictive modeling tools such as logistic regression or machine learning can further support stroke risk assessments at scale and help deliver timely, data-informed care to those who need it most.

By identifying the patients most at risk and supporting them earlier in life, healthcare providers can play a crucial role in reducing the long-term burden of stroke on individuals, families, and communities.

6.0 References

Soriano, F. (2021). *Stroke Prediction Dataset*. Retrieved from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Centers for Disease Control and Prevention (CDC). (2023). *Defining Adult Overweight & Obesity*. https://www.lb7.uscourts.gov/documents/19-927URL1defining.pdf

World Health Organization (WHO). (2022). *Hypertension*. https://www.who.int/news-room/fact-sheets/detail/hypertension

American Diabetes Association. (2023). *Diagnosis*. https://diabetes.org/diabetes/a1c/diagnosis

JMP. (2024). *Chi-Square Test*. SAS Institute Inc. Retrieved from https://www.jmp.com/en/statistics-knowledge-portal/chi-square-test

7.0 Links to Logs, Analysis, and Repository

Data Cleaning and EDA Log

Google Sheets - Stroke Risk Analysis Logs

Stroke Risk Analysis - Final Dataset & Visualizations

Google Sheets - Stroke Risk Analysis Final

Github Repository

Github - github.com/jmedinacs/stroke-risk-analysis