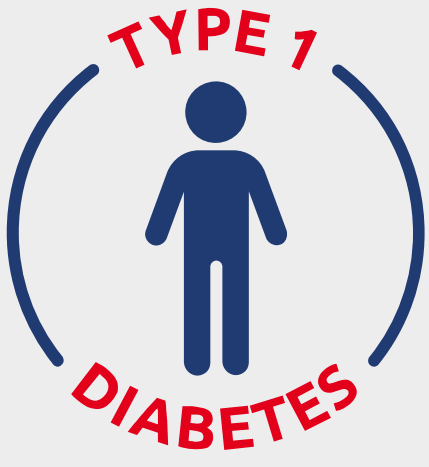
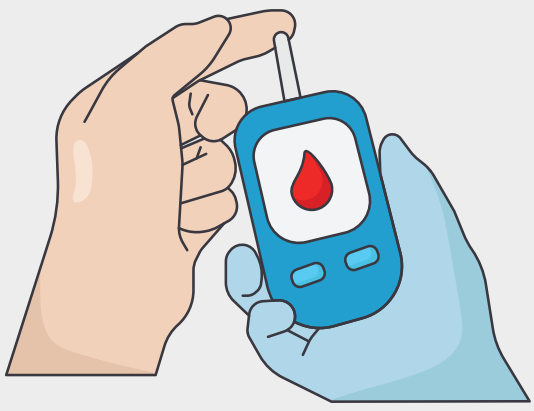
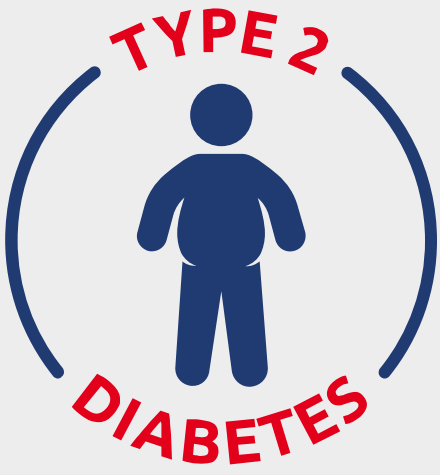




Understanding Diabetes Disease Risk Factors

By: Anna De Sanctis



Introduction

What is Diabetes?

Diabetes is a chronic disease that affects how the body turns food into energy. Normally, the body breaks down most of the food into glucose and releases it into the bloodstream. When blood sugar rises, the pancreas releases insulin, which helps cells absorb the sugar. In people with diabetes, this process does not work properly. Either the body does not make enough insulin, or it cannot use the insulin it produces effectively. Over time, this can lead to serious health complications including heart disease, kidney failure, nerve damage, vision loss, and stroke.

According to the Centers for Disease Control and Prevention, more than 38 million people in the United States have diabetes, which is about 11.6 percent of the population. An estimated 8.7 million of them are undiagnosed, meaning they are unaware they are living with the condition. Additionally, around 98 million adults in the U.S. have prediabetes, which puts them at a significantly higher risk of developing type 2 diabetes.

Diabetes is not only a major health concern but also a significant economic burden. The American Diabetes Association estimates that the total annual cost of diagnosed diabetes in the U.S. is 412.9 billion dollars. This figure includes direct medical expenses as well as reduced productivity caused by diabetes-related complications. People diagnosed with diabetes, on average, have medical costs that are 2.6 times higher than those without the disease.

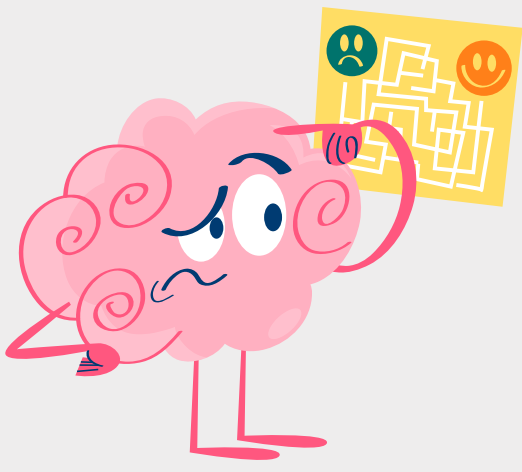
Though type 2 diabetes is often preventable through healthy lifestyle choices, millions of Americans remain at risk. The Cleveland Clinic notes that key risk factors include being overweight, physical inactivity, a family history of diabetes, high blood pressure, and high cholesterol. If not managed early, diabetes can silently cause damage to vital organs and blood vessels. In fact, complications such as heart attacks or kidney failure often emerge before a person even knows they are diabetic.

Who is at risk?

The National Institute of Diabetes and Digestive and Kidney Diseases reports that type 2 diabetes most often occurs in people over the age of 45, though it is increasingly common among young adults, teens, and children. In the United States, certain racial and ethnic groups including African Americans, Hispanic and Latino Americans, Native Americans, and some Asian American and Pacific Islander populations are at higher risk of developing type 2 diabetes. Other contributing factors include a lack of access to healthy food, limited healthcare options, poor living conditions, and low levels of physical activity. These are known as social determinants of health.

Recent research highlights that these non-clinical factors can be just as important as medical ones when identifying who is likely to develop diabetes. People living in communities with high poverty rates, low food access, or high levels of pollution often face compounded health risks. The CDC emphasizes that one in five people with diabetes do not even know they have it, making early identification especially important.

How to solve the problem?



The key to reducing the long-term damage and costs associated with diabetes lies in early detection. Identifying people at risk before symptoms appear can help reduce complications and prevent the disease from progressing. This can be done by understanding which factors most strongly predict the likelihood of someone having diabetes and using that knowledge to design a targeted screening system.

To build such a solution, data from multiple sources must be used. Demographic data such as age, household size, education level, and income provide an overview of someone's living conditions and access to resources. Personal health data, including pre-existing conditions like obesity, high blood pressure, and arthritis, help reveal physical risk factors. Finally, county-level data from organizations like the Robert Wood Johnson Foundation allow us to see how the environment and local healthcare systems influence disease outcomes.

By combining these data points, we can build a model that identifies who is most at risk of having undiagnosed diabetes. This predictive tool allows healthcare providers, public health agencies, and insurers to take action sooner, improving health outcomes and reducing long-term costs. While diabetes continues to affect millions of Americans, this model offers a path forward by using the data we have to protect the lives we can still change.

A Simple Solution

Before building a more complex model, I started with a basic question. What traits or conditions are most commonly associated with people who already have diabetes?

To explore this, I created a dependent variable using the available ailment data by identifying individuals who showed signs or symptoms related to diabetes. I then examined how individual variables like high blood pressure, obesity, and arthritis were associated with diabetes outcomes.

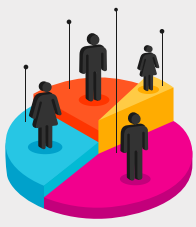
My analysis showed that high blood pressure and obesity had the strongest relationships with diabetes. These findings aligned with national data from the CDC and the American Diabetes Association. However, I quickly realized that using only one variable at a time was not enough. Some individuals with these conditions did not have diabetes, and others with diabetes did not have those specific conditions. This made it clear that simple one-to-one associations would leave too many cases undetected.

I also explored demographic traits such as income level, household size, and education. While certain groups showed slightly higher diabetes rates, no single demographic factor stood out as a strong individual predictor.

This early analysis showed me that diabetes risk is influenced by many overlapping factors. Although basic associations were helpful, they lacked the depth and accuracy needed to make meaningful predictions. To better identify individuals at risk, I decided to build a more complete model that included multiple variables. This led me to use regression analysis as the next step.

Persona Groups

Demographics



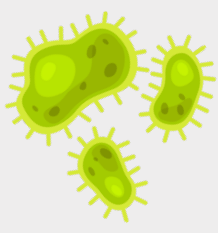
Mail Persona
- Most likely to order mail
- Most likely to donate to charities and politicians

Donation Persona
- Most likely to donate to animals and children
- Most likely to have a gold credit card

Income Persona
- Highest household income
- Highest net worth
- Highest education level

Household Persona
- Largest household size and # of children
- Least likely to have a gold or premium credit card

Ailments



Pain Persona
- Most likely to experience pain
- Most likely to have arthritis, insomnia, and asthma

Medicine Persona
- Most likely to take medicine for ailments
- Least likely t0 have a heart attack

Heart Condition Persona
- Most likely to have a heart condition, heart attack, or stroke
- Most likely to have high blood pressure

Cancer Persona
- Most likely to have cancer
- Most likely to be obese
- Least likely to have anxiety

County Health

Poor Health Areas
- Most Likely to have poor health
- Lowest Life Expectancy
- Most likely to smoke
- Most prevalent teen births

Hispanic Population Areas
- Highest Hispanic Population
- Largest Population
- Most likely to be uninsured

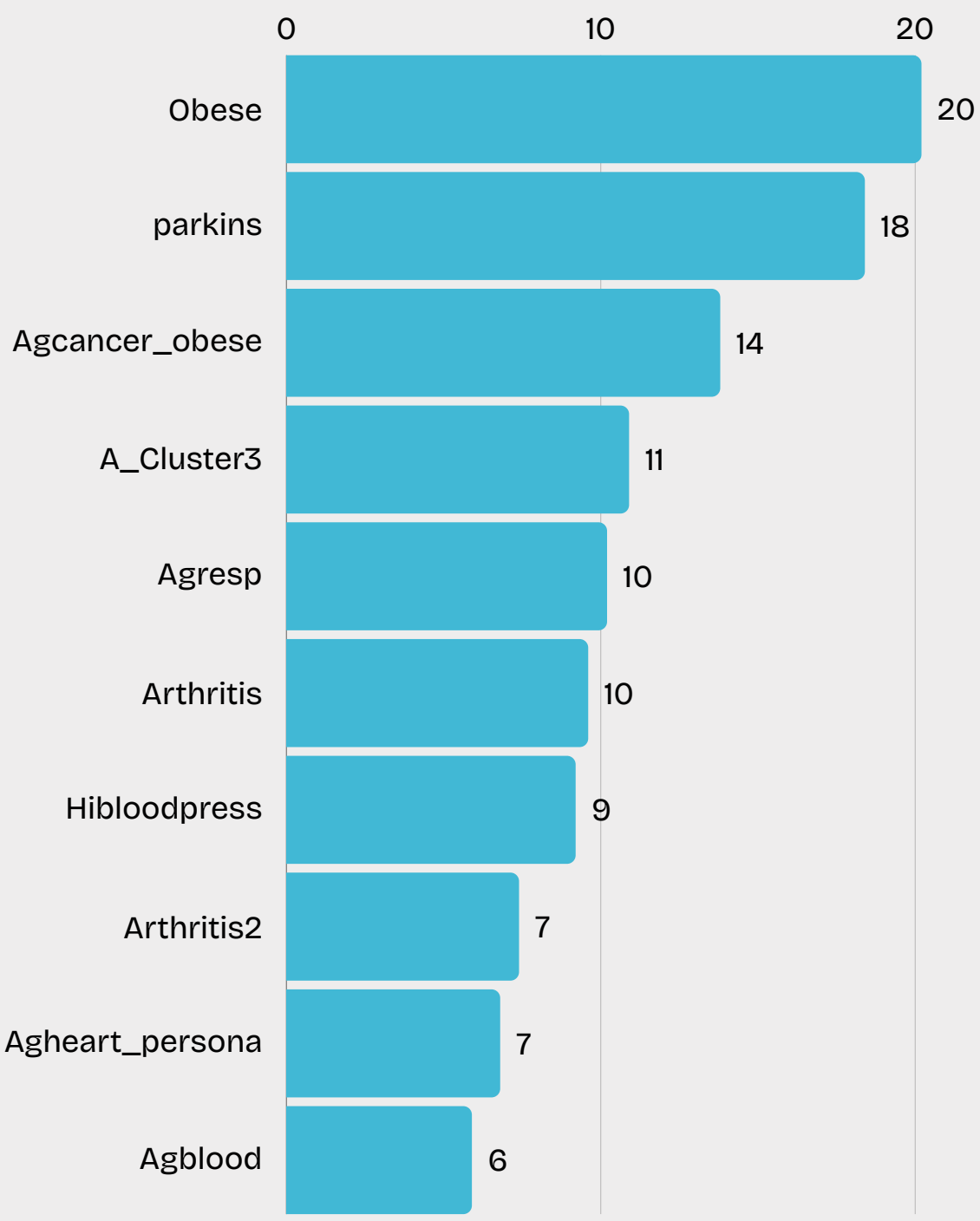
Life Expectancy Areas
- Most Likely to have HIV
- Lowest suicide rate
- Highest Life Expectancy

Violent Crime Areas
- Most likely to have violent crime
- Highest gun fatality rate
- Most likely to die due to drugs

Regression

When analyzed in isolation, personal ailment data provided the strongest predictive power for diabetes, outperforming demographic and county health data. However, combining all three sources — ailments, demographics, and personas — enabled a deeper and more accurate model. The final regression output achieved an R-squared of 0.280, meaning that the model explains 28 percent of the variation in diabetes diagnoses. This result demonstrates the value of layering different types of information to better understand health risk.

To visualize the relative importance of each variable, a bar chart of standardized coefficients was created. Obesity emerged as the strongest predictor with a weight of 20, followed closely by Parkinson's disease at 18 and obesity-related cancer conditions at 14. These results align with clinical findings that excess body weight and chronic conditions are highly correlated with type 2 diabetes.



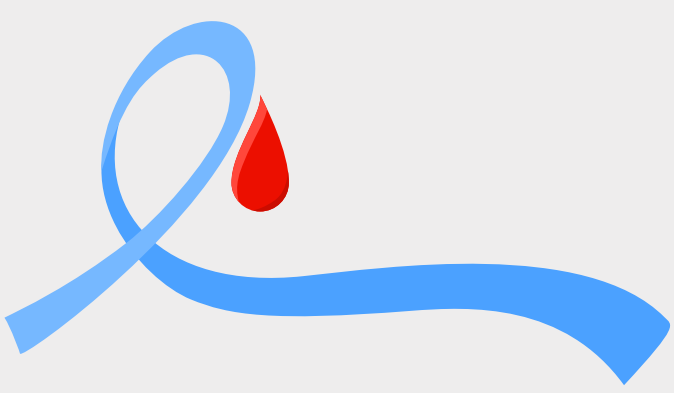
Each of these variables was statistically significant at the $p < .001$ level and reflects a distinct dimension of diabetes risk, from physical symptoms to underlying conditions. The bar chart confirms that obesity, neurological disorders, and inflammatory conditions are consistently associated with higher likelihoods of having diabetes.

This model not only identifies who is most at risk, but also helps health professionals and policymakers understand why certain people are at risk. The more data we incorporate—personal, medical, and social—the more precisely we can target interventions and preventative care. This insight is essential for reducing both the human and financial costs of undiagnosed diabetes in the United States.

Model Results

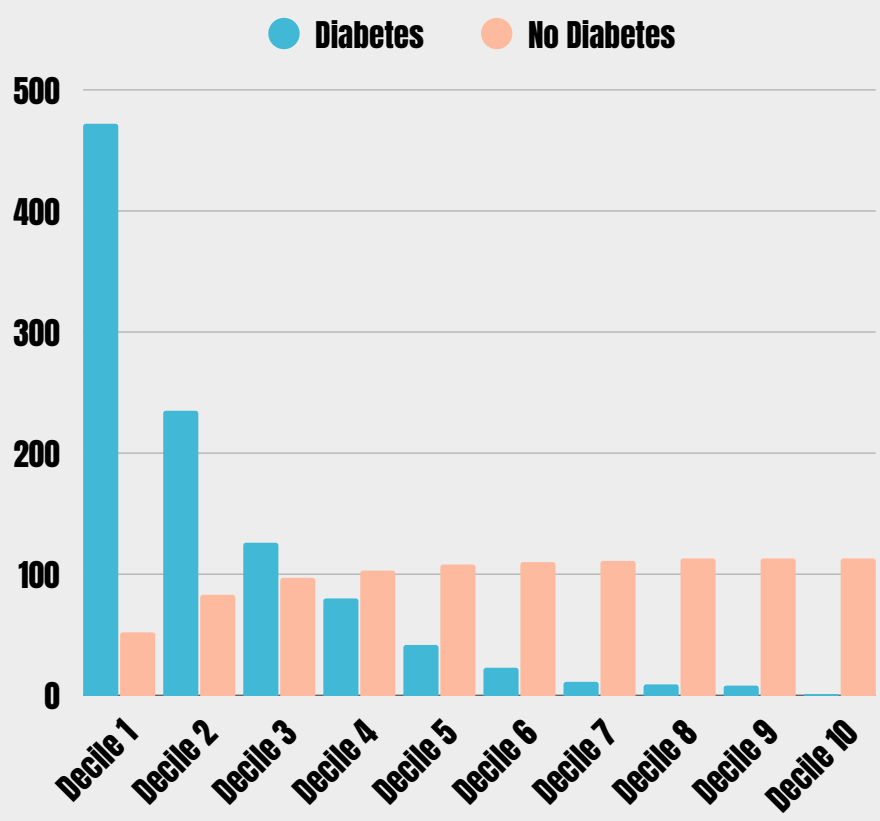
The regression model assigned a diabetes risk score to each person in the dataset. These scores were grouped into ten deciles, each representing ten percent of the population, ranked from highest to lowest predicted risk. Decile 1 contains the top ten percent with the highest scores, while Decile 10 includes those with the lowest.

The model performs very well at identifying people with diabetes. Nearly half of all diagnosed cases (47.2 percent) fall into the top decile. When expanding to the top two deciles, the model captures 70.7 percent of all known diabetes cases. This means that more than two out of every three people with diabetes appear in the top 20 percent of scores.



This level of performance suggests that the model is both accurate and efficient at flagging high-risk individuals. While there are false positives—people predicted to be at risk who are not currently diagnosed—the high concentration of known cases in the upper deciles supports the model's practical use for screening.

The chart below shows a steep drop in diabetes cases after Decile 2, with Deciles 3 through 10 each capturing fewer than 13 percent of the total diagnosed population. By Decile 10, only 0.1 percent of diagnosed individuals appear, reinforcing the idea that low scores are strongly associated with low diabetes risk.



These results are especially important in the context of undiagnosed diabetes. Since 23 percent of all diabetes cases in the U.S. are undiagnosed, it is reasonable to assume that some of the people with high scores but no current diagnosis may actually have diabetes but remain unaware. These individuals fall into the model's "false positive" category but could in fact be true positives awaiting diagnosis.

If healthcare providers used this model for targeted outreach or follow-up screenings, it could lead to earlier diagnoses and potentially prevent costly complications. The American Diabetes Association estimates that early treatment can save **$8,400** per person per year. In large populations, that translates to millions in annual savings and improved quality of life for patients.

Conclusion Insights

Diabetes remains one of the most underdiagnosed and expensive health issues in the United States, with an estimated 8.7 million people unaware they have it. My model addresses this gap by identifying the individuals most at risk. In my sample, over 70 percent of diagnosed cases appeared in the top two deciles, and more than 80 percent fell in the top three. This shows that the model is highly effective at concentrating risk, which can lead to faster and more focused screening.

If applied at scale, even modest use of this model could lead to significant results. For example, identifying just 87 undiagnosed individuals in a test population would save approximately 730,800 dollars annually in preventable healthcare costs, using the American Diabetes Association estimate of 8,400 dollars per person. Applied to a provider network like Kaiser Permanente, these savings could exceed 1.6 billion dollars per year.

This model is not only accurate but also actionable. It uses existing data that health systems already collect and transforms it into targeted interventions that can save lives and reduce financial pressure. For healthcare leaders, insurers, and policymakers, investing in this predictive tool offers a clear opportunity to improve outcomes and use resources more efficiently.

Bibliography

Chatterjee, R., Narayan, K. M. V., Lipscomb, J., and Phillips, L. S. "Screening Adults for Pre-Diabetes and Diabetes May Be Cost-Saving." Diabetes Care, vol. 33, no. 7, 2010, pp. 1484–1490. https://doi.org/10.2337/dc10-0054.

"Diabetes." Centers for Disease Control and Prevention, 15 May 2024, https://www.cdc.gov/diabetes/php/data-research/.

"Diabetes." Cleveland Clinic, https://my.clevelandclinic.org/health/diseases/7104-diabetes. Accessed 8 June 2025.

"Diabetes Statistics." National Institute of Diabetes and Digestive and Kidney Diseases, 12 Sept. 2024, https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics.

Emily D. Parker, et al. "Economic Costs of Diabetes in the U.S. in 2022." Diabetes Care, vol. 47, no. 1, 2 Jan. 2024, pp. 26–43. https://doi.org/10.2337/dci23-0085.

"Health and Economic Benefits of Diabetes Interventions." National Center for Chronic Disease Prevention and Health Promotion, 15 May 2024, https://www.cdc.gov/nccdphp/priorities/diabetes-interventions.html.

Khan, T., Yang, J., and Wozniak, G. "Trends in Medical Expenditures Prior to Diabetes Diagnosis: The Early Burden of Diabetes." Population Health Management, vol. 24, no. 1, 2021, pp. 46–51. https://doi.org/10.1089/pop.2019.0143.

"New American Diabetes Association Report Finds Annual Costs of Diabetes to Be $412.9 Billion." American Diabetes Association, 1 Nov. 2023, Add a little bit of body text.

"Population Clock." U.S. Census Bureau, Add a little bit of body text. Accessed 8 June 2025.

Warrington, J. S., et al. "Integrating Social Determinants of Health and Laboratory Data: A Pilot Study to Evaluate Co-Use of Opioids and Benzodiazepines." Academic Pathology, vol. 6, 2019, https://doi.org/10.1177/2374289519884877.

"Kaiser Foundation Health Plan and Hospitals Report 2023 Financial Results." Kaiser Permanente, 9 Feb. 2024, https://about.kaiserpermanente.org/news/press-release-archive/kaiser-foundation-health-plan-and-hospitals-report-2023-financial-results.