Comparative Analysis of Diabetes Risk in Pima and NHANES Populations

Laurentiu-Nicolae Fratila

May 28, 2025

Abstract

This report investigates and compares diabetes-related health indicators in two different populations: the Pima Indian cohort and a sample from the U.S. population based on the NHANES 2017–2018 survey. Statistical analyses were conducted on variables such as glucose levels, BMI, age, and diabetes prevalence.

1 Introduction

Diabetes is a chronic disease with rising prevalence worldwide. Certain populations, such as the Pima Indians, show particularly high risk. This study compares key health indicators between a dataset of Pima Indian women and a sample of U.S. women from NHANES.

2 Data and Methods

2.1 Pima Indian Diabetes Dataset

The Pima Indian dataset is a widely used public dataset originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of 768 female patients of Pima Indian heritage, all aged 21 or older. Each observation includes numeric variables such as number of pregnancies, plasma glucose concentration, diastolic blood pressure, skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age. The outcome variable is a binary indicator of whether the individual has diabetes (1) or not (0), based on diagnostic criteria at the time.

2.2 NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) is a program of the Centers for Disease Control and Prevention (CDC) that collects health-related data on a representative sample of the U.S. population. For this study, we used the 2017–2018 cycle and combined data from four components: demographic information, the diabetes questionnaire, fasting glucose lab values, and body measures. We restricted the sample to adult females aged 21 years and older, to align it with the Pima dataset.

2.3 Variable Harmonization

To allow comparison between the two datasets, we selected common variables: age, BMI, fasting plasma glucose, and diabetes status. In NHANES, diabetes status was derived from the self-reported variable `DIQ010`, with a value of 1 interpreted as a positive diabetes diagnosis. Gender was harmonized by selecting only female participants. Variables were renamed and cleaned to match those in the Pima dataset.

2.4 Statistical Analysis

All statistical analyses were performed in R. Descriptive statistics were computed for age, BMI, glucose, and diabetes prevalence in both datasets. We used two-sample t-tests to compare the mean values of continuous variables and a chi-square test to compare proportions of diabetes diagnosis between groups. Logistic regression models were fit separately on each dataset to assess the relationship between age, BMI, and glucose as predictors of diabetes status.

All plots were exported in Encapsulated PostScript (EPS) format to ensure compatibility with LaTeX.

3 Results

3.1 Descriptive Statistics

Table 1 summarizes the average values of age, BMI, and glucose in the two cohorts. On average, the NHANES sample is older but has lower BMI and glucose levels than the Pima Indian sample.

Table 1: Descriptive statistics for both datasets

Variable	Pima (mean)	NHANES (mean)	p-value
Age (years)	33.24	45.70	<0.001
BMI (kg/m^2)	31.99	28.96	<0.001
Glucose (mg/dL)	120.89	111.82	<0.001

3.2 Visual Comparison of Distributions

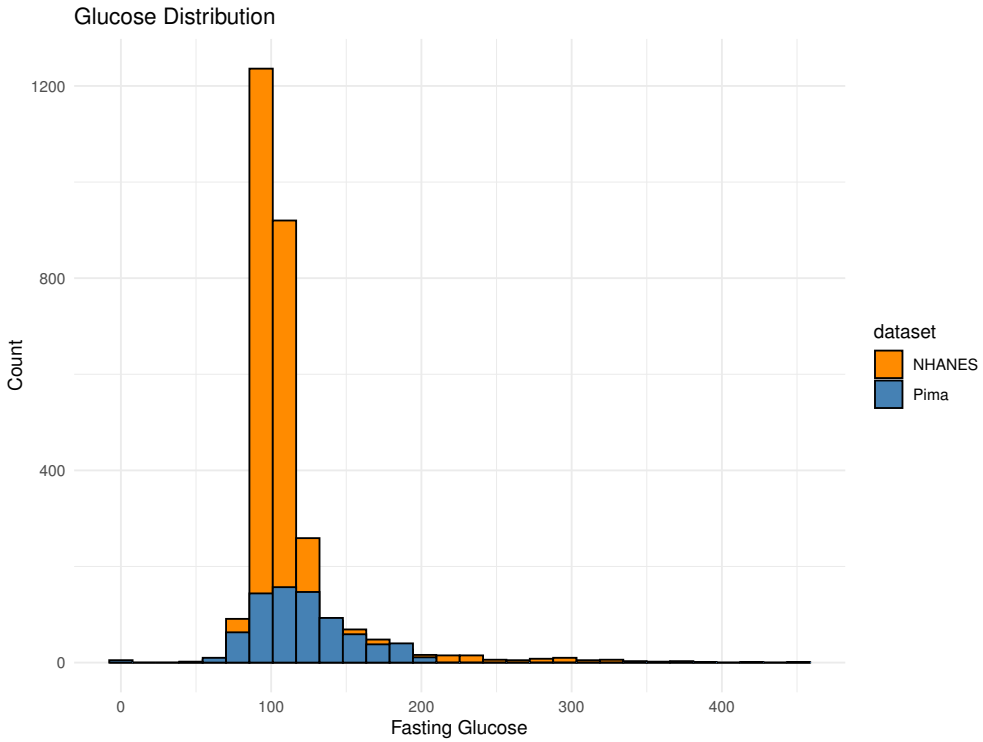

Figure 1: Distribution of fasting glucose levels in Pima and NHANES datasets.

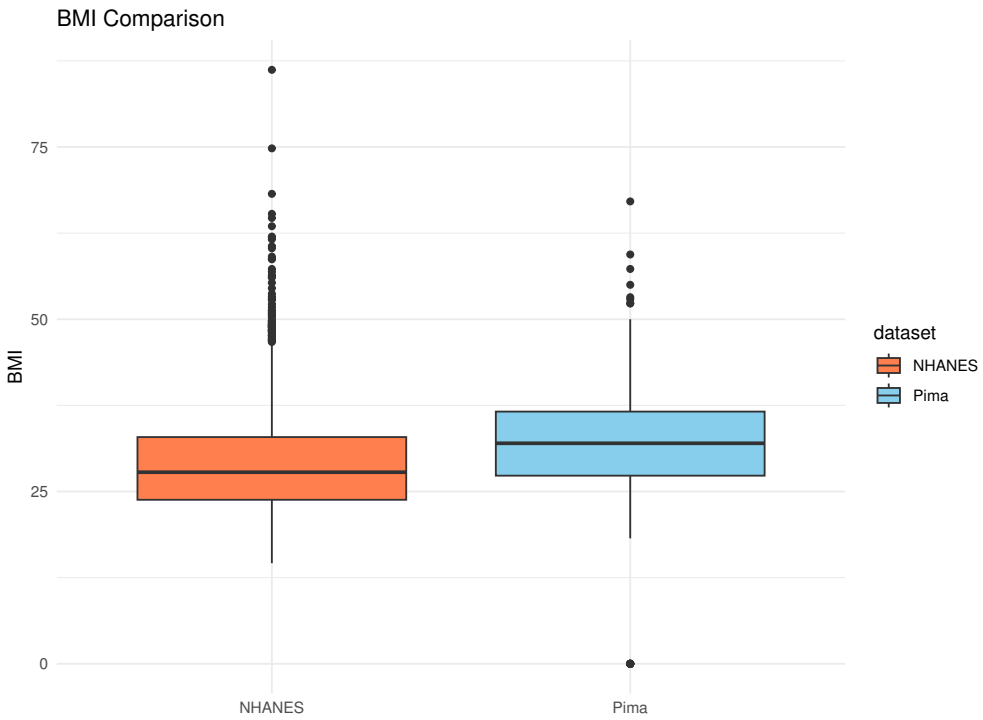

Figure 2: Comparison of BMI between Pima and NHANES cohorts.

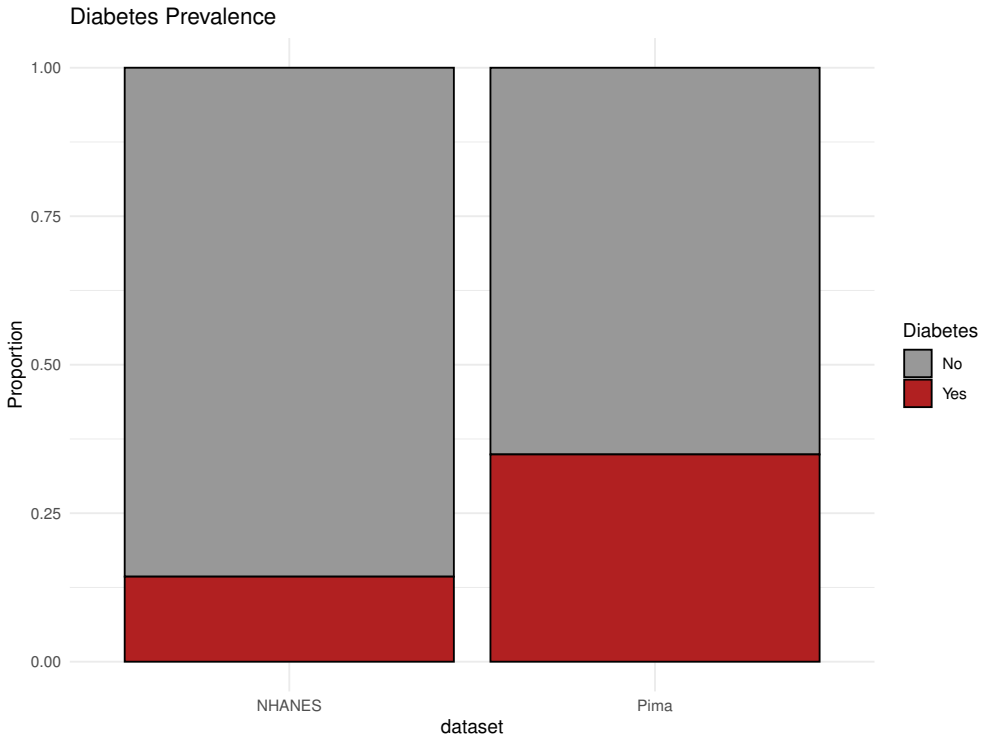

Figure 3: Diabetes prevalence in Pima vs NHANES groups.

3.3 Inferential Statistics

Two-sample t-tests confirmed significant differences in age, BMI, and glucose between the groups ($p < 0.001$ for all comparisons). A chi-square test for diabetes prevalence yielded a significant result ($\chi^2 = 167.02$, $df = 1$, $p < 0.001$), suggesting that diabetes rates differ significantly between the populations.

3.4 Logistic Regression Models

Separate logistic regression models were fitted on each dataset to predict diabetes status using age, BMI, and glucose as predictors. The results are summarized in Table 2.

Table 2: Logistic regression results: coefficients and p-values

Predictor	**Pima**		**NHANES**	
	Coef	p-value	Coef	p-value
Age	0.030	7.77e-05	0.058	<0.001
BMI	0.082	<0.001	0.054	<0.001
Glucose	0.033	<0.001	0.041	<0.001

4 Discussion and Conclusion

This study compared diabetes-related indicators between the Pima Indian cohort and a subset of the general U.S. population from the NHANES 2017–2018 survey. The Pima dataset, which includes only adult women of Pima Indian heritage, is known for a high prevalence of type 2 diabetes. The NHANES sample was filtered to include only adult women aged 21 and older, to match the characteristics of the Pima group as closely as possible.

Our findings reveal statistically significant differences in all key health indicators. The NHANES participants were, on average, older but had lower mean BMI and glucose levels. Diabetes prevalence was significantly higher in the Pima group compared to NHANES. These results are consistent with previous literature that identifies Native American populations as having elevated diabetes risk due to genetic, socioeconomic, and environmental factors.

Logistic regression models showed that age, BMI, and glucose levels were all significant predictors of diabetes status in both datasets. Glucose had the strongest predictive value, followed by BMI. The higher coefficient for age in the NHANES model likely reflects the wider age distribution and older mean age in that population.

This study highlights the importance of population-specific risk profiles in diabetes screening and prevention strategies. Although models trained on one population may be generalizable, they should be validated carefully before clinical application across diverse groups.

Limitations: The NHANES diabetes status is based on self-reported diagnosis, which may lead to underestimation. The Pima dataset is limited in scope and lacks some variables that NHANES includes, such as HbA1c levels. Furthermore, the study does not account for socioeconomic or lifestyle differences between the populations.

Conclusion: The Pima Indian cohort displays significantly higher diabetes prevalence and risk factor levels compared to a nationally representative female sample. This reinforces the need for targeted interventions and supports the broader use of biostatistical modeling for population health management.

References

1. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey (NHANES) 2017–2018. *Available at:* `https://www.cdc.gov/nchs/nhanes/index.htm`

2. National Center for Health Statistics. NHANES 2017–2018 Datasets. *Available at:* `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017`

3. UCI ML Repository. Pima Indians Diabetes Database. *Available at:* `https://www.kaggle.com/data`

4. American Diabetes Association. Standards of Medical Care in Diabetes—2024. *Diabetes Care*. 2024;47(Suppl 1):S1–S154. *Available at:* `https://diabetesjournals.org/care/articl`

5. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: Estimates for 2000 and projections for 2030. *Diabetes Care*. 2004;27(5):1047–1053.