Detailed Findings: Pima Indians Diabetes Risk Analysis

Executive Summary

This analysis examined 768 female patients from the Pima Indians Diabetes Database to identify key predictors and risk patterns for diabetes. The study revealed that **glucose levels, BMI, and age are the strongest predictors**, with risk compounding dramatically when multiple factors are present. Patients with all three major risk factors (age 40+, BMI 30+, glucose 126+) showed a **78.8% diabetes rate**, compared to just **7% for those with no risk factors** - an 11-fold increase in risk.

---

1. Dataset Overview and Population Characteristics

Basic Statistics

- **Total Patients:** 768 (all female, Pima Indian heritage)
- **Diabetes Prevalence:** 268 patients (34.9%) have diabetes
- **Non-Diabetic Patients:** 500 patients (65.1%)

Population Demographics

- **Average Age:** 33.2 years overall
    - Diabetic patients: 37.1 years
    - Non-diabetic patients: 31.2 years
- **Age Range:** 21-81 years
- **Age Distribution:** Heavily weighted toward younger patients (51.6% under 30)

---

2. Glucose Levels: The Primary Biological Marker

Average Glucose Comparison

- **Diabetic patients:** 141.3 mg/dL
- **Non-diabetic patients:** 109.0 mg/dL
- **Difference:** 32.3 mg/dL (29.6% higher in diabetics)

Clinical Significance

Glucose emerged as the single strongest predictor of diabetes, which aligns with clinical diagnostic criteria. Diabetes is diagnosed when fasting blood glucose exceeds 126 mg/dL, and our data confirms this threshold's validity.

Glucose Category Analysis

Normal Range (<100 mg/dL):

- 192 patients (25.2% of dataset)
- Only 14 have diabetes (7.3% diabetes rate)
- **Interpretation:** This represents the baseline risk level in the population

Pre-Diabetic Range (100-125 mg/dL):

- 274 patients (35.9% of dataset)
- 76 have diabetes (27.7% diabetes rate)
- **Interpretation:** Even in the "pre-diabetic" range, diabetes risk is nearly 4x higher than normal
- **Clinical implication:** Aggressive intervention needed at this stage

Diabetic Range (≥126 mg/dL):

- 297 patients (38.9% of dataset)
- 176 have diabetes (59.3% diabetes rate)
- **Interpretation:** More than half of patients with diabetic-level glucose actually have the diagnosis
- **Note:** 41% in this range do NOT have diagnosed diabetes, suggesting either:
    - Early stage disease not yet diagnosed
    - Recent glucose spike
    - Measurement timing issues

Percentile Analysis

Comparing the top 25% vs bottom 25% of glucose levels reveals the dramatic impact:

Bottom 25% (Glucose: 21-99 mg/dL):

- 191 patients
- 14 have diabetes (7.3% rate)
- Average glucose: 79.8 mg/dL

Top 25% (Glucose: 157-199 mg/dL):

- 191 patients
- 131 have diabetes (68.6% rate)
- Average glucose: 171.5 mg/dL

Key Finding: Being in the top glucose quartile increases diabetes risk by **9.4 times** compared to the bottom quartile. This is the largest single-factor difference found in the entire analysis.

---

3. Body Mass Index (BMI): The Modifiable Risk Factor

Average BMI Comparison

- **Diabetic patients:** 35.1 (Class 2 Obesity)
- **Non-diabetic patients:** 30.3 (Class 1 Obesity)
- **Difference:** 4.8 BMI points

Notable: Even the non-diabetic group averages in the obese category, reflecting the high obesity rates in the Pima Indian population.

BMI Category Breakdown

Underweight (BMI <18.5):

- 4 patients (0.5% of dataset)
- 0 have diabetes (0% diabetes rate)
- **Note:** Sample too small for meaningful conclusions

Normal Weight (BMI 18.5-24.9):

- 102 patients (13.4% of dataset)
- 7 have diabetes (6.9% diabetes rate)
- **Interpretation:** This represents the baseline risk for normal weight individuals

Overweight (BMI 25-29.9):

- 179 patients (23.5% of dataset)
- 40 have diabetes (22.3% diabetes rate)
- **Interpretation:** Risk more than triples compared to normal weight (22.3% vs 6.9%)

Obese (BMI ≥30):

- 472 patients (62% of dataset - majority of population!)
- 219 have diabetes (46.4% diabetes rate)
- **Interpretation:** Risk increases 6.7-fold compared to normal weight
- Nearly half of obese patients have diabetes

Clinical Implications

BMI is the **most actionable risk factor** because it's modifiable through lifestyle changes. The data suggests:

- Reducing BMI from obese to normal could reduce diabetes risk by approximately **85%** (46.4% → 6.9%)
- Even modest weight loss (obese → overweight) could cut risk in half (46.4% → 22.3%)

Why BMI Matters

Excess body fat, particularly abdominal fat, causes:

- Insulin resistance (cells don't respond properly to insulin)
- Chronic inflammation
- Increased free fatty acids interfering with glucose metabolism
- Hormonal imbalances affecting insulin production

---

4. Age Distribution and Risk Progression

Age Group Analysis

Under 30 (21-29 years):

- 396 patients (51.6% of dataset - largest group)
- 84 have diabetes (21.2% diabetes rate)
- **Interpretation:** Baseline/young adult risk level

Age 30-45:

- 254 patients (33.1% of dataset)
- 126 have diabetes (49.6% diabetes rate)
- **Critical Finding:** Risk more than DOUBLES at age 30
- **Interpretation:** This represents a critical intervention window

Age 46-60:

- 91 patients (11.8% of dataset)
- 51 have diabetes (56% diabetes rate - HIGHEST RISK GROUP)
- **Interpretation:** Peak diabetes risk period
- Over half of patients in this age range have diabetes

Over 60:

- 27 patients (3.5% of dataset - smallest group)
- 7 have diabetes (25.9% diabetes rate)
- **Interpretation:** Rate drops, likely due to:
    - Survivor bias (healthier individuals live longer)
    - Small sample size limiting reliability
    - Those with severe diabetes may not survive to 60+

The Age 30 Threshold

The most striking finding is the **dramatic jump at age 30** - from 21.2% to 49.6%. This suggests:

1. Cumulative lifestyle factors begin manifesting clinically around age 30
2. Metabolic changes associated with aging accelerate diabetes development

3. Years of obesity/poor diet reach a tipping point
4. Screening and prevention should intensify in late 20s

Why Age Matters

As people age:

- Beta cells in pancreas decline (less insulin production)
- Insulin sensitivity decreases
- Muscle mass decreases (muscle uses glucose)
- Physical activity often declines
- Years of dietary habits compound
- Hormonal changes affect metabolism

---

5. Multiple Risk Factors: Compounding Effects

Two-Factor Analysis: Obesity + Age Over 40

Both Risk Factors (Obese AND Over 40):

- 133 patients
- 80 have diabetes (60.2% rate)
- **Highest combined risk**

Obese BUT Under 40:

- 339 patients
- 139 have diabetes (41% rate)

Over 40 BUT Not Obese:

- 59 patients
- 21 have diabetes (35.6% rate)

Neither Risk Factor (Not Obese AND Under 40):

- 226 patients
- 26 have diabetes (11.5% rate)
- **Lowest risk baseline**

Key Finding: Having both risk factors creates a **5.2x higher risk** (60.2% vs 11.5%) compared to having neither. This demonstrates that risks don't just add - they multiply.

Three-Factor Risk Accumulation Analysis

This analysis examined patients with 0-3 major risk factors:

- Age ≥40
- BMI ≥30 (obese)
- Glucose ≥126 (diabetic range)

0 Risk Factors:

- 171 patients (22.5% of dataset)
- 12 have diabetes (7% rate)
- **Interpretation:** Near-baseline risk despite genetic predisposition

1 Risk Factor:

- 273 patients (35.9% of dataset)
- 68 have diabetes (24.9% rate)
- **3.6x increase from 0 factors**

2 Risk Factors:

- 228 patients (30% of dataset)
- 121 have diabetes (53.1% rate)
- **7.6x increase from 0 factors**
- **2.1x increase from 1 factor**

3 Risk Factors:

- 80 patients (10.5% of dataset)
- 63 have diabetes (78.8% rate - CRITICAL RISK)
- **11.3x increase from 0 factors**
- **1.5x increase from 2 factors**

Critical Insight: This is the most powerful finding in the entire analysis. Each additional risk factor doesn't just add to risk - it multiplies it. With all three factors present, nearly 4 out of 5 patients have diabetes.

Clinical Application:

- Patients with 0-1 factors: Standard monitoring
- Patients with 2 factors: Intensive screening and prevention
- Patients with 3 factors: Aggressive intervention required (78.8% already have or will develop diabetes)

---

6. Above-Average Risk Profile Analysis

Using subqueries, I identified patients with BOTH:

- Above-average glucose (>120.9 mg/dL)
- Above-average BMI (>31.9)

Results:

- 195 patients meet both criteria (25.6% of dataset)
- 128 have diabetes (65.6% rate)
- Nearly **2 out of 3** in this group have diabetes

Interpretation: Being above average in both key metrics creates a critical risk profile. These patients should be priority targets for:

- Immediate glucose screening
- Weight management programs
- Lifestyle intervention
- Close medical monitoring

---

7. Other Health Indicators

Insulin Levels

- **Diabetic patients:** 100.3 μU/mL
- **Non-diabetic patients:** 68.4 μU/mL
- **Difference:** 31.9 μU/mL (46.6% higher in diabetics)

Interpretation: Higher insulin in diabetics suggests **insulin resistance** - the body produces more insulin to compensate for cells not responding properly. This is characteristic of Type 2 diabetes, where the problem isn't insulin production (initially) but insulin effectiveness.

Data Quality Issue: 374 patients (48.7%) have zero insulin values, indicating missing data. This limits the reliability of insulin-based analysis.

Blood Pressure

- **Diabetic patients:** 70.8 mmHg (diastolic)
- **Non-diabetic patients:** 68.2 mmHg
- **Difference:** 2.6 mmHg (3.8% higher)

Interpretation: Blood pressure shows minimal difference between groups, making it a **weak predictor** of diabetes in this dataset. While diabetes and hypertension often co-occur, blood pressure alone doesn't effectively identify diabetes risk in this population.

Skin Thickness (Triceps Skinfold)

- **Diabetic patients:** 22.2 mm
- **Non-diabetic patients:** 19.7 mm
- **Difference:** 2.5 mm (12.7% higher)

Interpretation: Skin thickness estimates body fat percentage. The higher measurements in diabetics align with BMI findings - more body fat correlates with diabetes risk.

Data Quality Issue: 227 patients (29.6%) have zero values, indicating missing measurements.

Pregnancy History

- **Diabetic patients:** 4.9 pregnancies average
- **Non-diabetic patients:** 3.3 pregnancies average
- **Difference:** 1.6 pregnancies

Pregnancy & Age Interaction Analysis:

High Pregnancies (5+) + Older (35+):

- 203 patients
- 104 have diabetes (51.2% rate)

Low Pregnancies (<5) + Older (35+):

- 77 patients
- 38 have diabetes (49.4% rate)

High Pregnancies (5+) + Younger (<35):

- 73 patients
- 28 have diabetes (38.4% rate)

Low Pregnancies (<5) + Younger (<35):

- 415 patients
- 98 have diabetes (23.6% rate)

Key Finding: Age is more influential than pregnancy count. Both older groups show ~50% diabetes rates regardless of pregnancies. However, among younger women, high pregnancy count does increase risk (38.4% vs 23.6%).

Possible Mechanisms:

- Gestational diabetes history increases Type 2 diabetes risk
- Hormonal changes from multiple pregnancies affect insulin sensitivity
- Correlation with age (more pregnancies = older = higher risk)
- Weight gain during/after pregnancies

---

8. Data Quality Assessment

Missing Data (Represented as Zeros)

The dataset contains medically impossible zero values that indicate missing measurements:

Variable	Zero Values	% of Dataset	Impact
Insulin	374	48.7%	High - limits insulin analysis
Skin Thickness	227	29.6%	Moderate - affects body fat estimates
Blood Pressure	35	4.6%	Low - minimal impact
BMI	11	1.4%	Low - minimal impact
Glucose	5	0.7%	Minimal - excludes from analysis

Implications:

1. **Insulin data is unreliable** - nearly half missing means insulin-based conclusions should be treated cautiously
2. **Skin thickness is partially unreliable** - 30% missing limits its usefulness as a body fat indicator
3. **Glucose and BMI are reliable** - less than 2% missing, making them trustworthy for analysis
4. All analyses excluding these zero values still maintain large enough sample sizes for statistical validity

Data Quality Recommendations

For future studies:

- Ensure complete data collection, especially for insulin measurements
- If resources are limited, prioritize glucose and BMI measurements (highest predictive value)
- Consider multiple glucose measurements to account for daily variation
- Document reasons for missing values (equipment failure, patient refusal, etc.)

---

9. Clinical and Public Health Implications

For Healthcare Providers

1. Screening Priorities:

- **Highest Priority:** Women 30-60, BMI ≥30, with family history
- **Medium Priority:** Women under 30 with BMI ≥30 OR over 40 with any BMI
- **Standard Monitoring:** All others, but intensify if glucose enters pre-diabetic range

2. Prevention Strategies:

- **Weight Management:** Most impactful intervention - reducing BMI from obese to normal could prevent ~85% of diabetes cases
- **Age-Based Intervention:** Begin intensive prevention in late 20s, before the age 30 risk jump
- **Glucose Monitoring:** Regular screening for high-risk groups, quarterly for those with 2+ risk factors

3. Risk Communication:

- Use the "3 risk factors" framework with patients
- Emphasize compounding nature of risks (not additive, but multiplicative)
- Focus on modifiable factors (weight) while acknowledging non-modifiable ones (age, genetics)

For Public Health Policy

1. Population-Level Interventions: Given that 62% of the population is obese, population-wide interventions are needed:

- Community exercise programs
- Nutrition education initiatives
- Access to healthy, affordable food
- Built environment changes (walkability, parks, recreation facilities)

2. Targeted Programs:

- Workplace wellness programs focusing on 30-60 age group
- Postpartum diabetes prevention for women with gestational diabetes
- Cultural sensitivity in interventions (Pima Indians have genetic predisposition)

3. Early Detection:

- School-based screening programs
- Community health fairs targeting high-risk populations
- Integration of diabetes screening into routine primary care

Economic Considerations

Cost of Prevention vs Treatment:

- Weight loss interventions: ~$500-2,000 per person annually

- Diabetes treatment: ~$9,600 per person annually (direct medical costs)
- Diabetes complications: $15,000-30,000+ annually
- **ROI:** Every dollar spent on prevention saves $5-10 in treatment costs

Targeting High-Risk Groups: The "3 risk factors" group (78.8% diabetes rate, 80 patients) represents:

- 10.5% of population
- 23.5% of all diabetes cases (63 out of 268)
- Intensive intervention in this small group could prevent 60+ diabetes cases

---

10. Limitations and Future Research

Study Limitations

1. Population Specificity:

- All patients are female Pima Indians
- Pima Indians have one of the highest genetic predispositions to diabetes globally
- Findings may not generalize to:
    - Males
    - Other ethnic groups
    - Populations without genetic predisposition

2. Cross-Sectional Design:

- Cannot establish causation, only correlation
- Don't know if high BMI caused diabetes or if diabetes caused weight gain
- No temporal sequence of risk factor development

3. Sample Size Limitations:

- Over 60 age group has only 27 patients
- Underweight category has only 4 patients
- Some subgroup analyses have limited statistical power

4. Missing Data:

- 49% missing insulin data
- 30% missing skin thickness data
- May introduce selection bias if missingness is not random

5. Measurement Limitations:

- Single glucose measurement (may not reflect typical levels)
- No HbA1c data (3-month glucose average)
- No information on diabetes duration or severity

- No data on diabetes treatment or management

Recommendations for Future Research

1. Longitudinal Studies:

- Follow patients over time to establish causal relationships
- Track BMI changes and subsequent diabetes development
- Identify critical intervention windows

2. Intervention Studies:

- Test weight loss programs in high-risk groups
- Compare different intervention strategies
- Measure cost-effectiveness

3. Expanded Populations:

- Replicate analysis in diverse ethnic groups
- Include male patients
- Compare genetic vs lifestyle risk factors

4. Additional Variables:

- Include dietary data
- Physical activity levels
- Socioeconomic factors
- Healthcare access
- Medication use
- Family history details

5. Predictive Modeling:

- Build risk prediction models
- Validate models in external populations
- Develop clinical risk calculators
- Machine learning approaches for pattern identification

---

11. Conclusions

This analysis of 768 Pima Indian women revealed clear, actionable patterns in diabetes risk:

Primary Findings

1. **Glucose is the strongest biological predictor** (9.4x risk difference between top and bottom quartiles)

2. **BMI is the most modifiable risk factor** (6.7x risk difference between obese and normal weight)
3. **Age 30-60 represents peak risk period** (50-56% diabetes rates)
4. **Risk factors compound multiplicatively, not additively** (11.3x higher risk with 3 factors vs 0)
5. **78.8% of patients with all three major risk factors have diabetes** - this group requires immediate intervention

Most Actionable Insights

For Individual Risk Assessment:

- Count your risk factors (Age 40+, BMI 30+, Glucose 126+)
- 0-1 factors: Standard care
- 2 factors: Intensive monitoring needed
- 3 factors: Immediate aggressive intervention required

For Prevention:

- Weight management is the single most impactful intervention
- Intervention should begin before age 30
- Regular glucose screening essential for high-risk groups

For Healthcare Systems:

- Target the 10.5% with all three risk factors (prevents 23.5% of cases)
- Implement age-based screening intensification at 30
- Focus resources on weight management programs

Final Thoughts

The compounding nature of diabetes risk factors creates both a challenge and an opportunity. The challenge is that once multiple risk factors are present, diabetes risk becomes extremely high. The opportunity is that interventions targeting even one factor - particularly weight - can dramatically reduce risk.

In a population where 62% are obese, population-level interventions addressing obesity could prevent thousands of diabetes cases. The data clearly shows that diabetes is not inevitable, even in a genetically predisposed population, for those who maintain healthy weight and lifestyle factors.

Analysis completed: October 2025
Dataset: Pima Indians Diabetes Database (768 patients)
Methods: SQL-based statistical analysis using SQL Server