# AI Colab Group 1 - Clinical Data Science & Modeling

Yara Yaghi, Naod Dawit, Kareem Aly, & Indira Kuppa

# Proposed Research Question

How does kidney function, as measured by creatinine clearance (CCR) and blood urea nitrogen (BUN), relate to β-cell dysfunction and type 2 diabetes risk, and does this relationship differ by smoking status and alcohol consumption among U.S. adults?

# Potential Hypothesis

Reduced kidney function, indicated by lower creatinine clearance and higher BUN, is associated with impaired β-cell function and increased risk of type 2 diabetes, especially among current smokers and drinkers.

# Research Scope and Objectives

**Scope:**

**Population**: U.S. Adults from NHANES

**Key Explanatory Variables**:

- Kidney function: Creatinine Clearance (CCR), Blood Urea Nitrogen (BUN)
- Behavioral factors: Smoking status, alcohol consumption

**Mediating/Target Outcome:**

- β-cell dysfunction (estimated via HOMA-B)
- Type 2 diabetes status (diagnosed or based on biomarkers)

**Stratification/Interaction**: Smoking and alcohol status groups

**Objectives:**

**Primary**: Assess how kidney function markers (CCR and BUN) are associated with β-cell dysfunction and type 2 diabetes

**Secondary**: Support early identification of high-risk groups based on renal-endocrine interactions. Determine whether impaired renal function precedes or parallels endocrine dysfunction in T2D

# Research Potential

## Clinical Insight

- Improves early diabetes screening by flagging high-risk individuals with both metabolic red flags and family history
- Promotes personalized prevention strategies

## Research Contribution

- Explores an under-researched physiological pathway linking kidney and endocrine health
- Moves beyond traditional T2D predictors (e.g., BMI, glucose) toward organ-level systems modeling

## Public Health Relevance

- Offers insights for screening and risk stratification based on routine kidney function labs
- Informs prevention strategies for smokers and drinkers, who may be more vulnerable to combined renal and endocrine stress

# 1.Data Source - National Health and Nutrition Examination Survey)

Source: NHANES (National Health and Nutrition Examination Survey); large, nationally representative U.S. adult sample. Data collected through interviews, physical examinations, and laboratory tests.

- Dataset Size: ~30,000+ participants per combined multi-year cycles.

Target Outcomes → Binary Variables

- Diabetes Status: Self-reported diagnosis and laboratory-confirmed diabetes (Fasting Plasma Glucose ≥ 126 mg/dL or HbA1c ≥ 6.5%).

Predictors Included: Kidney Function:

- Blood Urea Nitrogen (BUN)
- Creatinine Clearance Rate (CCR)

β-cell Function:

- Fasting Insulin & Fasting Glucose (HOMA-β calculation)

Family History:

  - Family history of diabetes (parent, sibling) collected via questionnaire.

Physical Behaviors:

  - Smoking status, alcohol use, physical activity levels.

- Link to the Dataset -

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=1999

# Machine Learning Potential

**Predictive Modeling:**
Logistic Regression, Random Forest, XGBoost, LightGBM for type 2 diabetes classification.

**Moderation and Interaction Analysis:**
Machine learning to detect interactions between metabolic biomarkers, family history, and lifestyle.

**Risk Stratification:**
Identifying high-risk subgroups using combined NHANES and Kaggle datasets.

**Feature Importance & Explainability:**
SHAP (SHapley Additive exPlanations), feature ranking, and visualization to improve clinical interpretability.

**Unsupervised Learning:**
Clustering to explore hidden diabetes phenotypes.

# Web Scraping Integration

## Supplementary Data Collection:

- Scrape region-specific diabetes prevalence rates, lifestyle trends, and public health reports from CDC, WHO, and other health portals.
- Extract recent research abstracts and news updates from PubMed to track emerging diabetes risk factors.

## Added Value:

- Enriches model context with real-time, location-based, and literature-based insights.
- Allows integration of external environmental or policy-level factors.

## Tools:

- Python (BeautifulSoup, Requests, Pandas)
- Target websites: CDC Diabetes Atlas, BRFSS, PubMed

# Literature Review

**Metabolic & Behavioral factors:**

- Individuals with high-normal fasting glucose (91–99 mg/dL) have a greater risk of developing diabetes compared to those with lower normal levels (Brambilla et al., 2011)
- Triglycerides are an independent and early predictor of type 2 diabetes (Zhao et al., 2019)
- High BMI amplifies risk across all metabolic risk markers (Zhao et al., 2019)
- Smoking and alcohol use worsen metabolic regulation (Akhuemonkhan & Lazo, 2017)

# Literature Review

**Beta Cell Dysfunction**

- Beta cells located in the pancreas produce and secrete insulin (Dludla et al., 2023).
- Beta cell dysfunction indicates impaired insulin secretion, contributing to T2DM (Dludla et al., 2023).
- The Homeostatic Model Assessment of Beta-cell Function (HOMA-B) and the Insulinogenic Index can be used to indicate beta cell function (Kim et al., 2024; Sung et al., 2009)
    - These indicators can be calculated with fasting and post-load glucose and insulin values.

# Literature Review

**Type II Diabetes:**

- Type II diabetes, also known as adult-onset diabetes, occurs when the body is not able to utilize insulin correctly and sugar builds up in the blood (Mayo Clinic, 2025).
    - Type II diabetes is more common in older adults (hence adult-onset), however, more and more children are being diagnosed with the rise of childhood obesity (Mayo Clinic, 2025).
- As of 2024, more than 38 million Americans have diabetes, with close to 95% of diagnoses being for Type II diabetes (CDC, 2024)
    - Mostly in adults over 45 years old, but more and more children are getting diagnosed.

# References


Akhuemonkhan, E. & Lazo, M. (2017). Association between family history of diabetes and cardiovascular disease and lifestyle risk factors in the U.S. population (NHANES 2009–2012). *Preventive Medicine, 96*, 129–134. https://doi.org/10.1016/j.ypmed.2016.12.015

Baliunas, A., Taylor, B., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., & Rehm, J. (2009). Alcohol as a Risk Factor for Type 2 Diabetes. *Diabetes Care, 32*(11), 2123-2132. https://doi.org/10.2337/dc09-0227

Brambilla, P., La Valle, E., Falbo, R., Limonta, G., Signorini, S., Cappellini, F., & Mocarelli, P. (2011). Normal fasting plasma glucose and risk of type 2 diabetes. *Diabetes Care, 34*(6), 1372–1374. https://doi.org/10.2337/dc10-2263

CDC. (2024, May 15). About Type 2 Diabetes. Diabetes. https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html

Dludla, P. V., Mabhida, S. E., Ziqubu, K., Nkambule, B. B., Mazibuko-Mbeje, S. E., Hanser, S., Basson, A. K., Pheiffer, C., & Kengne, A. P. (2023). Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. *World Journal of Diabetes, 14*(3), 130–146. https://doi.org/10.4239/wjd.v14.i3.130

Kim, J. Y., Lee, J., Kim, S. G., & Kim, N. H. (2024). Recent Glycemia Is a Major Determinant of β-Cell Function in Type 2 Diabetes Mellitus. *Diabetes & Metabolism Journal*, *48*(6), 1135-1146. https://doi.org/10.4093/dmj.2023.0359

Mayo Clinic. (2025, February 27). Type 2 diabetes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Sung, K. C., Reaven, G., & Kim, S. (2009). Utility of Homeostasis Model Assessment of β-Cell Function in Predicting Diabetes in 12,924 Healthy Koreans. *Diabetes Care, 33*(1), 200-202. https://doi.org/10.2337/dc09-1070

Zhao, J., Zhang, Y., Wei, F., Song, J., Li, W.-D., Chen, C., Zhang, K., & Feng, S. (2019). Triglyceride is an independent predictor of type 2 diabetes among middle-aged and older adults: A prospective study with 8-year follow-ups in two cohorts. *Journal of Translational Medicine, 17*, 354. https://doi.org/10.1186/s12967-019-02156-3

# DATA FILES AND SOURCES

https://www.cdc.gov/brfss/annual_data/annual_2015.html

https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system/data

https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data