Predicting Diabetes: Identifying Key Physical and Socioeconomic Health Factors

Phoenix Nénar Williams

December 5, 2025

1 Introduction

According to the Centers for Disease Control and Prevention, approximately 14.7% of U.S. adults had diabetes in 2021, while an additional 38.0% met clinical criteria for prediabetes based on fasting glucose and HbA1c levels [1]. Despite this prevalence, an estimated 22.8% of adults with diabetes were unaware of their condition, and only 19% of prediabetic individuals reported being informed of their status by a healthcare provider.

Rates of undiagnosed or unrecognized diabetes are disproportionately higher among women, racial and ethnic minority populations, and older adults [1]. As a result, diabetes remains both highly prevalent and frequently underdetected, complicating prevention, treatment, and disease management efforts. Moreover, heterogeneity in disease etiology and clinical presentation makes it difficult for patients and caregivers to identify which health factors are most critical to address.

This study examines how demographic, metabolic, lifestyle, and socioeconomic factors are associated with diabetes risk across multiple disease types and diagnosis statuses, with the goal of identifying predictors that are both stable and clinically interpretable.

2 Biochemistry of Diabetes

Diabetes occurs when insulin production is insufficient (Type 1 or gestational diabetes) or when insulin signaling is impaired due to insulin resistance (Type 2 diabetes) [2, 7]. When a person eats food, the digestive system passes the sugars in the food into the blood. In the blood, insulin bonds to those sugars, transporting them safely to other places in the body to convert them into energy. However, if the body does not produce enough insulin or struggles to bond insulin and sugars correctly, the sugars can get trapped in the bloodstream. In a chemical process called glycation, these sugars form strong bonds with other proteins, lipids, and nucleic acids, inhibiting those molecules from executing other functions they are designed to perform. Over time and in extreme cases, the glycated proteins, lipids, and nucleic acids can make the blood viscous enough to impede blood flow as well. Both of these issues can put those with diabetes at higher risk for a plethora of other health issues, many of which can be fatal. These biochemical processes increase the risk of cardiovascular disease, organ damage, and metabolic dysfunction [7]. Diabetes remains uncurable, but treatable.

3 Diagnosis of Diabetes

Diabetes or prediabetes is most often diagnosed by fasting glucose, insulin, or blood sugar levels (see Table 1 for expected ranges), but identifying patients who should be treated for diabetes in the first place can be difficult because the symptoms may be undetected for long stretches of time before being noticed. Importantly, the struggle to recognize symptoms and risk or diagnose diabetes is linked to socioeconomic circumstances and healthcare access in addition to the body's ability to hide the symptoms. Depending on those levels and other health information, a patient can be diagnosed with any of the following types of diabetes.:

3.1 Type 1

Type 1 diabetes is characterized by the body ceasing to produce insulin, forcing those with it to take insulin every day to survive. The prevailing theory suggests that it is caused by an autoimmune reaction, but other causes are possible. Because of the lack of clarity pertaining to its causes and its quick onset,

Table 1: Diabetic Ranges of Blood Work Measurements

Index	Non-Diabetic	Prediabetic	Diabetic
Fasting Glucose	84.36 – 100.68	118.39 – 121.73	106.91 – 235.59
Postprandial Glucose	113.47 – 131.69	182.77 – 190.31	186.68 – 338.16
Glycated Hemoglobin (HbA1c)	4.35 – 5.37	5.67 – 5.95	6.61 – 8.23
Total Cholesterol	139.58 – 170.92	206.44 – 240.84	252.37 – 321.81
Triglycerides	88.34 – 126.12	118.60 – 184.88	159.95 – 279.59
Low-Density Lipoprotein (LDL)	71.34 – 102.88	135.28 – 169.10	182.39 – 253.33
High-Density Lipoprotein (HDL)	41.89 – 51.49	38.03 – 44.17	21.27 – 29.89

it is generally unpreventable and unpredictable. Generally, those with Type 1 diabetes are aware of their condition from a young age, but healthcare access, socioeconomic circumstances, and the possibility for it to be caused by pancreas failure means that can be diagnosed at any age. Only affecting 5-10% of all people with diabetes, it is rarer than other types. Because of its characteristics, it will generally present with very low insulin alongside the high blood sugar, fasting glucose, and post-meal glucose levels.

3.2 Type 2

Type 2 diabetes is characterized by the body failing to use insulin properly and uncontrollable blood sugar levels. Contrary to Type 1, type 2 is preventable, as it is caused by the body building up resistance to insulin over time. Thus, the age of diagnosis and the treatment for type 2 diabetes may differ from patient to patient, depending upon the risk factors and clinical measurements that are present in the individual patient. Because of its association with lifestyle, it is often stigmatized more than other types, even though it is the most common type. Contrary to Type 1 diabetes, high insulin levels may be present in Type 2 patients alongside the high blood sugar and glucose levels.

3.3 Gestational

While gestational diabetes is characterized by the same insulin resistance as type 2 diabetes, it only affects pregnant people and can subside post-pregnancy. It is generally asymptomatic but poses significant health risks for the pregnant person and fetus, so most pregnant people are tested for gestational diabetes at around 24-28 weeks, when it usually develops. The need to treat the pregnancy at the same time as diabetes often complicates its treatment plan from the norms for Type 2 patients, partially due to pregnant people needing to gain weight to support the health of the fetus rather than lose it. The clinical measurements for insulin, blood sugar, and glucose may be similar to what is found in Type 2, but because of the complications that pregnancy induces, it is far less stigmatized. Importantly, gestational diabetes can increase the pregnant person's risk for recurrence in future pregnancies as well as increase the pregnant person's and child's risk for type 2 diabetes later in life.

3.4 Prediabetes

Prediabetes is a health classification designed by medical professionals to attempt to prevent the progression to Type 2 diabetes. Specifically, prediabetes is a range for clinical measurements of insulin, blood sugar, and glucose that bridges the gap between healthy and diabetic levels. People whose blood work falls into the prediabetic range can often prevent or halt the progression of their diabetes with key lifestyle

changes. The challenges of diagnosing and treating prediabetes make it all the more essential to increase awareness of prevention methods and support for those who do have it.

4 The Diabetes Health Indicators Dataset

The observational dataset used in this analysis was obtained from a publicly available Kaggle repository containing simulated health indicators for diabetes research [4]. Importantly, location information is absent from this dataset and all demographic information is very general, substantially decreasing the risk of identifying patients from the data provided. Of the 31 variables, 12 are best described as factor type variables and the other 19 are numeric type variables (6 integer type, 13 float type variables). Table 2 displays the variable names, types, descriptions, values, and category for each variable. The source for the dataset did not indicate a specific data collector, but did suggest using it for classification models, regression models, experimental data analysis, machine learning, and hypothesis testing, suggesting that it was collected for practice with various mathematical concepts at least as much as it was collected for studying the health factors that affect diabetes risk. The dataset creator also provided Table 3 to interpret the risk score in the dataset.

4.1 Dataset Construction and Initial Cleaning

To address the heterogeneity in both the causes and clinical presentation of diabetes, the data were stratified by diabetes type and diagnosis status prior to modeling.

Specifically, the dataset was split by the categorical variables `type` and `diagnosis`, yielding up to ten possible combinations. Of these, seven resulted in nonempty datasets:

- No Diabetes, Undiagnosed
- Pre-Diabetes, Undiagnosed
- Type 1 Diabetes, Undiagnosed
- Type 1 Diabetes, Diagnosed
- Type 2 Diabetes, Diagnosed
- Gestational Diabetes, Undiagnosed
- Gestational Diabetes, Diagnosed

Three combinations contained zero observations: No Diabetes Diagnosed, Pre-Diabetes Diagnosed, and Type 2 Diabetes Undiagnosed. The absence of diagnosed cases among individuals without diabetes or with prediabetes is consistent with clinical definitions. However, the complete absence of undiagnosed Type 2 diabetes cases is noteworthy, particularly given the known prevalence of undiagnosed Type 2 diabetes in the general population. This suggests that the dataset may overrepresent clinically identified Type 2 diabetes cases, potentially due to the data collection or simulation process.

After splitting, each dataset was subjected to initial data cleaning procedures. Observations with missing values in the response variable `risk` were removed. Predictor variables were retained in their original scale to preserve interpretability, and all models were restricted to numerical variables to control model complexity and ensure stability across datasets with smaller sample sizes.

5 Results

5.1 Exploratory Data Analysis

Exploratory analysis was conducted separately for each of the seven datasets to assess variable distributions, detect anomalies, and evaluate the suitability of predictors for linear regression modeling.

Across all datasets, numeric variables exhibited varying degrees of skewness, particularly for blood biomarkers such as triglycerides, insulin, fasting glucose, and postprandial glucose. These variables generally displayed right-skewed distributions, reflecting the presence of individuals with extreme but clinically plausible measurements. Given the clinical interpretability of these variables and the focus on prediction rather than strict normality of predictors, no transformations were applied at this stage.

Sample sizes varied considerably across datasets. The largest dataset, Type 2 Diabetes Diagnosed, contained over 59,000 observations prior to filtering, whereas the smallest datasets—Type 1 Diabetes Undiagnosed and Gestational Diabetes Undiagnosed—contained fewer than 100 observations. These differences informed later decisions regarding model complexity, influence diagnostics, and interpretation of results. The exploratory analysis suggested that while the datasets were generally well-behaved, careful attention would be required to address multicollinearity among clinically related predictors and to assess the influence of extreme observations. These considerations guided the next stage of the analysis: collinearity assessment and variable selection.

5.2 Collinearity Assessment

Before proceeding to model selection, we assessed multicollinearity among the numeric predictors in each dataset to ensure coefficient stability, interpretability, and valid inference. Collinearity was evaluated using three complementary approaches: pairwise correlation matrices and scatterplot matrices, variance inflation factors (VIFs), and eigenvalue–condition proportion diagnostics. All assessments were conducted after initial data cleaning and variable screening, but prior to variable selection.

Across all datasets, collinearity patterns were broadly consistent, with the strongest dependencies arising among blood lipid measures and body composition variables. Importantly, no dataset exhibited multicollinearity severe enough to invalidate regression modeling after modest variable reduction.

5.2.1 Gestational Diabetes

Diagnosed. The gestational diagnosed dataset exhibited mild to moderate correlations among lipid-related variables. LDL cholesterol was positively correlated with triglycerides and negatively correlated with HDL cholesterol, while BMI showed moderate positive associations with triglycerides and insulin. Lifestyle variables such as sleep duration, screen time, and physical activity demonstrated weak correlations with both metabolic and demographic variables.

Variance inflation factors were uniformly below conventional thresholds of concern (VIF < 5 for all retained predictors), indicating that linear dependencies among predictors were not excessive. Eigenvalue proportion diagnostics similarly revealed no single dimension dominating variance across multiple predictors. As a result, no variables were removed at this stage due solely to collinearity.

Undiagnosed. The undiagnosed gestational dataset displayed nearly identical collinearity structure to the diagnosed group, though with slightly attenuated correlations among metabolic markers. Pairwise plots showed diffuse scatter without pronounced linear structure for most lifestyle variables. Lipid variables again formed the most correlated cluster, but VIFs remained modest and eigenvalue diagnostics

showed no evidence of near-singular design matrices. Consequently, all candidate predictors were retained for subsequent selection procedures.

5.2.2 No Diabetes (Undiagnosed)

In the no diabetes undiagnosed dataset, overall correlations were weaker than in diabetic cohorts. Demographic variables such as age exhibited small positive correlations with LDL cholesterol and triglycerides, while alcohol use and physical activity showed minimal association with most metabolic measures.

Collinearity diagnostics confirmed this visual impression. VIF values were close to one for the majority of predictors, and eigenvalue proportions were evenly distributed across dimensions. This dataset exhibited the least collinearity of all groups analyzed, requiring no variable exclusion at this stage.

5.2.3 Pre-Diabetes (Undiagnosed)

The pre-diabetes undiagnosed dataset showed modest strengthening of correlations relative to the non-diabetic group, particularly among BMI, triglycerides, LDL cholesterol, and insulin. HDL cholesterol retained a moderate negative association with LDL cholesterol, consistent with established physiological relationships.

Although several lipid variables exhibited statistically significant pairwise correlations, variance inflation factors remained below conservative cutoffs, and eigenvalue diagnostics did not suggest harmful redundancy. Given the clinical relevance of these variables and the absence of severe collinearity, all predictors were retained for model selection.

5.2.4 Type 1 Diabetes

Diagnosed. Among diagnosed Type 1 patients, correlations among blood lipid variables and insulin were present but weaker than those observed in Type 2 diabetes. BMI showed modest associations with triglycerides and LDL cholesterol, while lifestyle variables again displayed weak linear relationships with metabolic outcomes.

Collinearity diagnostics were favorable: VIF values were uniformly low, and eigenvalue proportions indicated no problematic accumulation of variance within a small number of components. This reflects the more heterogeneous metabolic presentation of Type 1 diabetes compared to insulin resistance–driven conditions.

Undiagnosed. The undiagnosed Type 1 dataset exhibited similar patterns, though with slightly noisier scatterplots due to smaller sample size. Correlation magnitudes were generally small, and no predictor pair approached thresholds typically associated with unstable coefficient estimates. As with the diagnosed group, all predictors were retained for downstream modeling.

5.2.5 Type 2 Diabetes (Diagnosed)

The diagnosed Type 2 diabetes dataset demonstrated the strongest collinearity structure among all groups analyzed. BMI, triglycerides, LDL cholesterol, and insulin formed a clearly correlated cluster, reflecting the central role of insulin resistance and dyslipidemia in Type 2 diabetes. HDL cholesterol showed a consistent negative association with LDL cholesterol.

Despite these physiologically expected relationships, collinearity diagnostics remained within acceptable bounds. While some VIF values were elevated relative to other datasets, none exceeded thresholds

warranting automatic exclusion. Eigenvalue proportion diagnostics suggested moderate shared variance across metabolic predictors but did not indicate near-linear dependence. Given the clinical interpretability of these variables and their importance to diabetes risk, collinearity was addressed primarily through variable selection rather than preemptive removal.

5.2.6 Summary of Collinearity Findings

Across all datasets, multicollinearity was present primarily among blood lipid and body composition variables, consistent with known metabolic relationships. However, no dataset exhibited collinearity severe enough to preclude regression analysis. Instead of aggressively removing correlated predictors at this stage, we retained clinically meaningful variables and relied on subsequent variable selection procedures to balance predictive performance, interpretability, and model stability.

Overall, the collinearity assessment supported the feasibility of fitting multiple linear regression models for diabetes risk across all disease categories while maintaining valid inference and coefficient interpretation.

5.3 Variable Selection

Variable selection was conducted separately for each dataset using a combination of backward elimination based on p-values, forward and stepwise selection using Akaike Information Criterion (AIC), and exhaustive subset selection evaluated using AIC, Bayesian Information Criterion (BIC), adjusted R^2, root mean squared error (RMSE), and Mallows' C_p. Emphasis was placed on identifying predictors that were consistently retained across multiple criteria, indicating robustness to model specification and selection approach.

5.3.1 Gestational Diabetes

Diagnosed. For the gestational diabetes diagnosed cohort, variable selection procedures consistently identified age, physical activity, HDL cholesterol, triglycerides, and postprandial glucose as key predictors of risk. These variables were retained across backward elimination, AIC-based forward and stepwise selection, and AIC-optimal subset models. More parsimonious criteria such as BIC favored a reduced model consisting primarily of age, physical activity, and HDL cholesterol, indicating that while metabolic variables improve predictive performance, a smaller core set explains much of the variability in risk. Variables such as alcohol use and insulin were selected only under adjusted R^2 and RMSE criteria, suggesting limited incremental explanatory value beyond the core predictors.

Undiagnosed. In the gestational diabetes undiagnosed group, age, body mass index (BMI), physical activity, HDL cholesterol, triglycerides, and diet score emerged as consistently important predictors. These variables were retained across backward elimination and all AIC-based procedures, while BIC again selected a more parsimonious subset emphasizing age, BMI, HDL cholesterol, and physical activity. Screen time appeared in models optimized for predictive accuracy but was not consistently retained across inferential criteria, indicating that behavioral variables contribute modestly to prediction but less so to stable explanatory structure.

5.3.2 No Diabetes (Undiagnosed)

Among individuals without diabetes and undiagnosed, variable selection revealed a broader set of influential predictors. Age, diet score, physical activity, HDL cholesterol, triglycerides, insulin, and LDL cholesterol were consistently retained across backward elimination, AIC-based methods, and subset selection optimized by adjusted R^2. BIC-selected models excluded glucose postprandial levels, suggesting that glycemic measures contribute less explanatory power in metabolically healthy populations. Variables such as screen time and sleep were primarily selected under RMSE and C_p criteria, reflecting their role in improving prediction rather than driving core risk associations.

5.3.3 Pre-Diabetes (Undiagnosed)

For the pre-diabetes undiagnosed cohort, selection results closely mirrored those observed in the non-diabetic group, with age, diet score, physical activity, HDL cholesterol, triglycerides, insulin, and LDL cholesterol consistently retained. The agreement between AIC- and BIC-based subset selection indicates a stable intermediate metabolic risk profile, where both lifestyle and biochemical factors contribute meaningfully to risk. Heart rate appeared only in adjusted R^2-optimized models, suggesting limited robustness as a primary explanatory variable.

5.3.4 Type 1 Diabetes

Diagnosed. In the type 1 diabetes diagnosed cohort, variable selection procedures favored a comparatively parsimonious model. Age, physical activity, insulin, and diet score were consistently selected across backward elimination, AIC-based approaches, and subset selection. BIC and Mallows' C_p criteria further emphasized age and physical activity as the most stable predictors. Alcohol use appeared intermittently in larger models but was not retained under stricter penalization, indicating limited independent contribution once insulin management and lifestyle factors were accounted for.

Undiagnosed. For individuals with undiagnosed type 1 diabetes, age, physical activity, insulin, screen time, and triglycerides were consistently selected across backward elimination and AIC-based procedures. BIC-based subset selection favored a slightly reduced model excluding triglycerides, while adjusted R^2 and RMSE criteria supported their inclusion. The emergence of screen time in multiple selection methods suggests a potential behavioral component to risk in undiagnosed cases, though its role appears secondary relative to metabolic and activity-related variables.

5.3.5 Type 2 Diabetes (Diagnosed)

The type 2 diabetes diagnosed cohort exhibited the most complex selection structure, reflecting substantial metabolic heterogeneity. Age, physical activity, HDL cholesterol, triglycerides, diet score, postprandial glucose, insulin, and LDL cholesterol were consistently retained across backward elimination, AIC-based methods, and exhaustive subset selection. More conservative BIC models excluded sleep and alcohol use, indicating that while behavioral variables may improve predictive performance, core metabolic and lifestyle factors dominate explanatory power in this population.

5.3.6 Summary

Overall, variable selection results demonstrate a clear progression from lifestyle-dominant predictors in metabolically healthy and gestational cohorts toward increasingly complex metabolic profiles in diagnosed

diabetes populations. Predictors retained across multiple selection criteria were prioritized in subsequent modeling stages to ensure interpretability, stability, and robustness of inference.

5.4 Influential Observations - Before Observation Removal

After finalizing candidate models through variable selection and accounting for model uniqueness, influential observation diagnostics were conducted for each dataset–model combination. Four standard influence measures were examined: high leverage points (based on hat values), Cook's distance, externally studentized residuals, and Bonferroni-adjusted outlier tests. These diagnostics were used to assess the stability of model estimates and identify observations that may disproportionately affect inference. Because the datasets vary substantially in size and heterogeneity, influence results were interpreted relative to dataset structure rather than absolute counts alone.

5.4.1 Gestational Diabetes

Diagnosed. For the gestational diabetes diagnosed cohort, influence diagnostics indicated a modest presence of influential observations. Backward p-value and subset-based models identified a moderate number of high-leverage points, particularly under the Subsets_Cp criterion, while Cook's distance and studentized residuals flagged fewer observations overall. Bonferroni-adjusted tests identified only a single extreme outlier across all model specifications. While some observations exert structural leverage in this cohort, the overall influence burden is limited and unlikely to undermine model stability.

Undiagnosed. In the gestational diabetes undiagnosed dataset, influence diagnostics revealed no high-leverage observations across all unique model specifications. However, Cook's distance and externally studentized residuals consistently flagged between eight and eleven observations as potentially influential. Despite this, Bonferroni-adjusted tests identified only one extreme outlier. The absence of leverage issues combined with modest influence measures suggests that the models for this cohort are relatively stable, with influence driven by localized residual behavior rather than structural imbalance.

5.4.2 No Diabetes (Undiagnosed)

The no diabetes undiagnosed cohort exhibited a substantial number of influential observations, particularly with respect to leverage. Both backward elimination and subset-based BIC models identified tens of thousands of high-leverage points, reflecting the large sample size and high-dimensional predictor space. Cook's distance and studentized residuals flagged several hundred observations across models, while Bonferroni tests identified a small number of extreme outliers. These findings indicate that influence in this cohort is driven primarily by structural heterogeneity rather than isolated anomalous cases.

5.4.3 Pre-Diabetes (Undiagnosed)

Influence diagnostics for the pre-diabetes undiagnosed cohort revealed the most pronounced influence patterns of all datasets analyzed. High-leverage observations numbered in the hundreds of thousands to over one million for models selected via backward elimination and adjusted R^2 criteria. Cook's distance and studentized residuals flagged several thousand observations, though Bonferroni-adjusted tests consistently identified only a single extreme outlier. These results reflect substantial heterogeneity within the pre-diabetic population and underscore the need for careful sensitivity analysis when interpreting model estimates for this group.

5.4.4 Type 1 Diabetes

Diagnosed. For individuals with diagnosed type 1 diabetes, influence diagnostics indicated minimal concern. Across all unique model specifications, at most two high-leverage observations were identified, with three to four observations exceeding Cook's distance or studentized residual thresholds. Only one observation was flagged by Bonferroni-adjusted testing. The limited presence of influential observations suggests a stable modeling structure and reliable inference for this cohort.

Undiagnosed. The undiagnosed type 1 diabetes cohort exhibited slightly elevated influence relative to the diagnosed group, though still at low absolute levels. No high-leverage observations were detected across any model specification, while four to six observations exceeded Cook's distance or studentized residual thresholds. A single extreme outlier was identified by Bonferroni adjustment. These findings suggest moderate localized influence without evidence of structural instability.

5.4.5 Type 2 Diabetes (Diagnosed)

Influence diagnostics for the diagnosed type 2 diabetes cohort revealed extensive leverage and influence, consistent with its large size and metabolic heterogeneity. Millions of high-leverage observations were identified under backward elimination and subset-based BIC and adjusted R^2 models. Cook's distance flagged between approximately 1,700 and 2,300 observations across models, while studentized residuals identified several hundred. Despite this, Bonferroni-adjusted tests again isolated only a single extreme outlier. These results indicate that influence is driven largely by scale and complexity rather than isolated pathological observations.

5.4.6 Summary

Across all datasets, influential observation diagnostics revealed a clear relationship between dataset size, heterogeneity, and the prevalence of leverage and influence. Smaller and more homogeneous cohorts, such as type 1 diabetes groups, exhibited minimal influence concerns, while larger populations—particularly pre-diabetes and type 2 diabetes cohorts—displayed substantial leverage and influence driven by structural complexity. Importantly, Bonferroni-adjusted outlier tests consistently identified very few extreme observations, suggesting that influence issues are predominantly systematic rather than driven by isolated anomalies. These findings motivated subsequent analyses comparing models fitted with and without influential observations to assess robustness.

5.5 Cook's D Filtering and Validation

To address the presence of influential observations identified in the initial diagnostics, Cook's distance was used to iteratively remove observations exceeding the threshold $4/(n-p)$, where n is the sample size and p is the number of model parameters. This procedure was applied separately to each unique model retained after variable selection. After filtering, each cleaned model was re-estimated and evaluated to ensure that (i) no influential observations remained under either Cook's distance threshold, and (ii) the cleaned dataset satisfied basic sample size requirements relative to model complexity. Models violating the $n \geq 10p$ rule after filtering were flagged as unstable and treated cautiously in subsequent analyses.

Gestational Diabetes. Both gestational diabetes datasets exhibited substantial sensitivity to influential observations, particularly for higher-complexity models. In the undiagnosed subset, Cook's distance

filtering removed between 18% and 58% of observations, and models selected by RMSE and adjusted R^2 became unstable due to insufficient post-filtering sample size. In contrast, simpler models selected by backward or forward p-value criteria remained stable after filtering. Among diagnosed gestational cases, most models retained stability, although the RMSE-based model violated the $n \geq 10p$ guideline and was excluded from further interpretation.

Pre-Diabetes (Undiagnosed). For undiagnosed pre-diabetes, Cook's distance filtering consistently removed approximately one-quarter of the dataset across most models. Despite the substantial reduction in sample size, all retained models satisfied post-filtering diagnostic criteria and exhibited no remaining influential observations. This stability suggests that influential points were primarily isolated rather than structurally embedded within the predictor space.

No Diabetes (Undiagnosed). In the no-diabetes undiagnosed dataset, Cook's distance filtering removed between 16% and 18% of observations depending on model specification. All models remained stable after filtering, reflecting the large baseline sample size and relatively moderate model complexity. These results indicate that influential observations were present but did not fundamentally compromise model reliability.

Type 1 Diabetes. Both diagnosed and undiagnosed type 1 diabetes datasets demonstrated high sensitivity to model complexity due to limited sample size. While several models remained stable after filtering, RMSE-optimized and certain subset-based models violated the $n \geq 10p$ rule and were flagged as unstable. These findings underscore the limitations of aggressive variable selection in small clinical subpopulations.

Type 2 Diabetes (Diagnosed). The diagnosed type 2 diabetes dataset experienced the largest absolute number of influential observations removed, with up to 39% of cases excluded in some models. However, due to the large initial sample size, all filtered models remained stable and free of remaining influential observations. This pattern suggests heterogeneity within the population rather than model misspecification.

Summary. Overall, Cook's distance filtering substantially improved model robustness across datasets by eliminating influential observations without inducing instability in most cases. Instability was confined primarily to small-sample datasets and high-complexity models, particularly those optimized for RMSE. Subsequent error assumption checks and prediction analyses were therefore conducted using Cook's-distance-filtered data for stable models, while unstable models were interpreted cautiously or excluded from downstream inference.

5.6 Post–Filtering Influence Diagnostics

After applying Cook's distance filtering and validating the stability of each retained model, influence diagnostics were recomputed to confirm that no influential observations remained according to Cook's distance thresholds. This step serves as a final verification that influential points were successfully mitigated and that remaining deviations are attributable to residual variation rather than leverage or undue influence. The diagnostics considered include high-leverage points, Cook's distance exceedances, large studentized residuals, and Bonferroni outlier tests.

Gestational Diabetes. For both diagnosed and undiagnosed gestational diabetes datasets, post-filtering diagnostics confirm that Cook's distance successfully eliminated influential observations. No models exhibited large Cook's distance values after filtering. A small number of high-leverage points and moderate numbers of large studentized residuals remained, which is expected in clinical datasets with heterogeneous physiological responses. These residual deviations did not correspond to Bonferroni-significant outliers and therefore do not indicate remaining influential observations.

No Diabetes and Pre-Diabetes (Undiagnosed). The undiagnosed no-diabetes and pre-diabetes datasets retained relatively large numbers of studentized residuals despite the absence of influential observations. This pattern reflects natural variability in risk scores within non-diabetic and pre-diabetic populations rather than model instability. Importantly, no Cook's distance exceedances were detected, confirming that influential points were effectively removed and that residual variation is not driven by leverage effects.

Type 1 Diabetes. Both diagnosed and undiagnosed type 1 diabetes datasets exhibited minimal post-filtering influence concerns. The small number of high-leverage points and large residuals observed is consistent with limited sample sizes and does not indicate undue influence. All retained models satisfied post-filtering influence criteria.

Type 2 Diabetes (Diagnosed). Although the diagnosed type 2 diabetes dataset displayed large numbers of studentized residuals, no influential observations remained after filtering according to Cook's distance. Given the large sample size and clinical heterogeneity of this population, such residual patterns are expected and do not undermine model validity. The absence of Cook's distance exceedances confirms that the final models are robust to influential observations.

Summary. Overall, post–Cook's distance diagnostics confirm the effectiveness of the filtering procedure across all datasets. While large studentized residuals persist in several large-sample datasets, no models exhibit remaining influential observations. This distinction reinforces the importance of separating residual variability from influence and supports the use of the Cook's-distance-filtered datasets for subsequent error assumption checks and prediction analyses.

5.7 Error Assumption Diagnostics

Following variable selection, influence filtering, and model validation, standard linear model error assumptions were assessed for each retained model. These diagnostics focused on (i) homoscedasticity via variance ratio F-tests, (ii) normality of residuals via the Shapiro–Wilk test, and (iii) independence of errors using Durbin–Watson statistics and lag-1 residual correlations. All diagnostics were conducted on Cook's-distance-filtered datasets, with original (unfiltered) results retained for comparison where relevant.

Overall, Cook's distance filtering substantially improved adherence to model assumptions, particularly with respect to variance homogeneity and serial independence. Deviations from normality remained common in larger datasets, consistent with the known sensitivity of normality tests to sample size. Importantly, no dataset exhibited systematic violations of independence, and heteroscedasticity was generally limited to a small subset of predictors rather than being pervasive across models.

5.7.1 Gestational Diabetes

Diagnosed. For gestational diabetes cases with prior diagnosis, variance ratio tests indicated mostly stable residual variance across predictors following filtering. Isolated deviations were observed for glucose postprandial levels and HDL cholesterol in some models, though these effects were not consistent across selection methods. Shapiro–Wilk tests frequently rejected normality, particularly for backward and subset-based models; however, these deviations were also present in the original data and are attributable to skewed clinical distributions rather than model misspecification. Durbin–Watson statistics remained close to 2 across all models, and lag-1 residual correlations were small in magnitude, indicating no meaningful autocorrelation.

Undiagnosed. In the undiagnosed gestational diabetes cohort, variance homogeneity improved markedly after Cook's distance filtering, with most predictors exhibiting non-significant variance ratio tests. Some variance instability persisted in smaller subset-based models, particularly those optimized for RMSE, reflecting reduced sample sizes rather than structural heteroscedasticity. Residual normality was generally acceptable for cleaned models, while strong non-normality was observed in the original data. Independence diagnostics consistently supported the absence of serial correlation.

5.7.2 No Diabetes (Undiagnosed)

The undiagnosed non-diabetic cohort demonstrated strong adherence to variance homogeneity assumptions across all cleaned models, with variance ratio statistics tightly centered around unity. Normality tests were not computed for this dataset due to large sample size, where trivial deviations would be expected to produce highly significant results. Durbin–Watson statistics were near 2 and lag-1 correlations were effectively zero, confirming independence of residuals. Collectively, these results indicate excellent error structure stability in this population.

5.7.3 Pre-Diabetes (Undiagnosed)

For undiagnosed pre-diabetes, variance ratio tests revealed mild heteroscedasticity for select metabolic predictors, including HDL cholesterol and triglycerides, particularly in RMSE-optimized models. These effects were statistically detectable but modest in magnitude and expected given metabolic heterogeneity in pre-diabetic populations. Residual normality tests were omitted due to large sample size. Independence diagnostics again showed no evidence of serial correlation, supporting the validity of the linear modeling framework.

5.7.4 Type 1 Diabetes

Both diagnosed and undiagnosed type 1 diabetes datasets exhibited generally acceptable variance homogeneity after filtering, with only sporadic deviations for insulin and triglycerides in smaller models. Residual normality tests frequently rejected normality in cleaned datasets; however, these deviations were also present in the original data and reflect physiological variability rather than modeling artifacts. Durbin–Watson statistics and lag-1 correlations provided no evidence of autocorrelation.

5.7.5 Type 2 Diabetes (Diagnosed)

In the diagnosed type 2 diabetes cohort, variance ratio tests were largely stable across predictors, though alcohol use and physical activity exhibited statistically detectable variance differences in some models.

Given the large sample size, these findings reflect minor departures rather than substantive heteroscedasticity. Normality tests were not performed due to sample size considerations. Residual independence was strongly supported, with Durbin–Watson statistics near 2 and negligible lag-1 correlations.

5.7.6 Summary

Across all datasets, Cook's distance filtering substantially improved error assumption compliance without introducing systematic violations. Residual independence was consistently satisfied, and variance heterogeneity was limited, predictor-specific, and clinically interpretable. Deviations from normality were most pronounced in large samples and are not expected to meaningfully affect inference. These diagnostics support the adequacy of the linear modeling framework for subsequent interpretation and comparison across disease states.

5.8 Interpretation of Regression Coefficients

The final regression models provide insight into how demographic, metabolic, and lifestyle variables are associated with estimated diabetes risk across different clinical subgroups. Coefficients are interpreted as the expected change in the risk score associated with a one-unit increase in the predictor, holding all other variables constant.

Because multiple model selection procedures were used, emphasis is placed on predictors that appear consistently across final models within each dataset and whose estimated effects are both statistically significant and substantively meaningful. Particular attention is given to the direction, magnitude, and stability of coefficients across diagnosed versus undiagnosed populations.

Table 10 summarizes the most consistently retained predictors and their qualitative effects across datasets.

Across all datasets, age and physical activity display the most consistent and interpretable effects, with age positively associated with risk and physical activity strongly protective.

5.8.1 Gestational Diabetes

Diagnosed. In the gestational diagnosed population, age exhibited a consistently positive association with risk, with estimates indicating that each additional year of age increased the risk score by approximately 0.3–0.5 units. Physical activity demonstrated a strong negative association, suggesting a substantial protective effect even after diagnosis.

HDL cholesterol was negatively associated with risk, aligning with its established cardiometabolic protective role. Glucose and triglyceride measures were occasionally retained but often failed to achieve statistical significance, suggesting that behavioral and lipid variables may better explain residual variation once diagnosis has occurred.

Undiagnosed For undiagnosed gestational cases, coefficient magnitudes were remarkably stable across models. Age and body mass index showed strong positive effects, while diet score and physical activity exhibited large negative coefficients, indicating that healthier behaviors were associated with substantially lower estimated risk.

Notably, physical activity coefficients were among the largest in absolute magnitude across all datasets, underscoring its importance as a modifiable risk factor in this group.

5.8.2 No Diabetes (Undiagnosed)

Among individuals without diagnosed diabetes, the models revealed a clear metabolic and behavioral gradient of risk. Age, triglycerides, insulin, and screen time were positively associated with risk, while HDL and physical activity were strongly protective.

The relatively small standard errors and consistent coefficient signs reflect the large sample size and suggest that even modest lifestyle changes may correspond to meaningful differences in estimated risk at the population level.

5.8.3 Pre-Diabetes (Undiagnosed)

The pre-diabetic undiagnosed group exhibited some of the strongest and most uniform coefficient patterns. Age, insulin, LDL cholesterol, and triglycerides were all positively associated with risk, while diet score and physical activity showed pronounced negative effects.

The magnitude of the physical activity coefficient was especially notable, indicating that increased activity levels are associated with large reductions in predicted risk even in individuals already exhibiting metabolic dysregulation.

5.8.4 Type 1 Diabetes

Diagnosed. Coefficient interpretation in the diagnosed Type 1 group was less stable. Although age remained positively associated with risk and physical activity retained a protective effect, many metabolic variables failed to reach statistical significance.

This instability likely reflects heterogeneity in disease management, insulin therapy, and clinical history, which introduce variability not fully captured by the available predictors.

Undiagnosed. In contrast, the undiagnosed Type 1 group demonstrated clearer and more consistent associations. Age, triglycerides, insulin, and screen time were positively associated with risk, while HDL, diet score, and physical activity were negatively associated.

Prior to diagnosis, risk in this population is strongly structured by observable metabolic and behavioral factors.

5.8.5 Type 2 Diabetes (Diagnosed)

For individuals with diagnosed Type 2 diabetes, coefficient estimates were highly stable and consistent across models. Age, insulin, triglycerides, LDL cholesterol, and screen time all showed strong positive associations with risk, while diet score, HDL, and physical activity were robustly protective.

The persistence of these effects even after diagnosis highlights the continued importance of lifestyle and metabolic management in controlling disease severity.

5.8.6 Summary

Across all datasets, several overarching patterns emerge. Age is universally associated with increased risk, reflecting cumulative metabolic burden over time. Physical activity consistently exhibits one of the largest protective effects, often exceeding that of individual metabolic markers in magnitude.

Lipid measures, particularly HDL and triglycerides, play a central role in distinguishing risk profiles across both diagnosed and undiagnosed populations. Finally, models fitted to undiagnosed groups tend

to yield more stable and interpretable coefficients, suggesting that treatment and disease management introduce additional complexity beyond what is captured by standard covariates.

Overall, the coefficient estimates reinforce the central role of modifiable lifestyle factors in diabetes risk and progression, even after accounting for demographic and biochemical characteristics.

5.9 Prediction

After completing variable selection and model diagnostics, the final stage of the analysis focused on prediction and interpretation. For each dataset corresponding to a diabetes status group, multiple candidate regression models were retained based on different selection criteria, including backward and forward p-value procedures as well as subset selection methods optimized for adjusted R^2, Bayesian Information Criterion (BIC), Mallows' C_p, and root mean squared error (RMSE).

To assess predictive performance in a consistent and interpretable manner, fitted values were computed at representative covariate profiles, along with both confidence intervals (CI) for the mean response and prediction intervals (PI) for individual-level outcomes. Confidence intervals quantify uncertainty in the estimated mean risk score, whereas prediction intervals reflect both estimation uncertainty and inherent variability in individual observations. Together, these measures provide a comprehensive picture of model reliability and clinical interpretability.

Table 11 summarizes representative predicted risk values across datasets, along with corresponding confidence and prediction intervals derived from the preferred subset-based models.

Across all datasets, confidence intervals are notably narrow for large-sample groups such as the undiagnosed and Type 2 populations, while prediction intervals remain substantially wider, reflecting meaningful inter-individual variability in risk.

5.9.1 Gestational Diabetes

Diagnosed. For individuals diagnosed with gestational diabetes, the best-performing models consistently identified age, high-density lipoprotein (HDL), and physical activity as key predictors of risk. Subset selection models, particularly those optimized for RMSE, achieved strong explanatory power ($R^2 \approx 0.94$), indicating that a combination of lifestyle and metabolic variables captures most of the variation in risk scores. The prediction intervals remain moderately wide, highlighting heterogeneity within this clinically complex population.

Undiagnosed. In the undiagnosed gestational group, predictive accuracy was exceptionally high, with adjusted R^2 values approaching unity for subset-based models. Risk increased strongly with age and body mass index, while higher diet quality and physical activity were protective. The extremely narrow confidence and prediction intervals suggest a highly stable risk structure in this subgroup, likely driven by large sample size and consistent behavioral patterns.

5.9.2 No Diabetes (Undiagnosed)

Among individuals without diagnosed diabetes, models again demonstrated very high explanatory power. Risk was strongly associated with age, triglycerides, screen time, and insulin levels, while HDL and physical activity exerted protective effects. Although the mean predicted risk was moderate, the prediction intervals indicate nontrivial individual-level variability, underscoring the presence of subclinical risk even in ostensibly healthy populations.

5.9.3 Pre-Diabetes (Undiagnosed)

The pre-diabetic undiagnosed group exhibited some of the strongest and most consistent predictor effects across all datasets. Age, lipid measures, insulin, and lifestyle variables contributed significantly, and all retained models achieved adjusted R^2 values exceeding 0.96. Prediction intervals, while wider than confidence intervals, remained relatively compact, reflecting both strong signal and large sample size.

5.9.4 Type 1 Diabetes

Diagnosed. Predictive performance for diagnosed Type 1 diabetes was notably weaker than for other datasets. Although age and physical activity remained significant predictors, overall model fit was modest ($R^2 \approx 0.30$–0.43). Wide prediction intervals reflect substantial heterogeneity, likely due to disease management differences, treatment regimens, and smaller sample size.

Undiagnosed. In contrast, the undiagnosed Type 1 group exhibited substantially stronger model performance, with subset RMSE models achieving adjusted R^2 values above 0.94. Risk was strongly associated with age, triglycerides, screen time, and physical activity. The comparatively narrow prediction intervals suggest more homogeneous risk patterns prior to formal diagnosis.

5.9.5 Type 2 Diabetes (Diagnosed)

Type 2 diabetes models demonstrated excellent predictive accuracy, with adjusted R^2 consistently above 0.96. Age, diet score, lipid profiles, insulin, and physical activity all played significant roles. Despite the strong model fit, prediction intervals remained meaningfully wide, reflecting real-world variability in disease severity and management.

5.9.6 Overall Interpretation and Implications

The prediction results reveal several consistent themes. Across nearly all datasets, age and physical activity emerged as robust predictors of risk, while lipid measures and dietary indicators played increasingly important roles in pre-diabetic and Type 2 populations. Models fitted to undiagnosed groups generally exhibited higher explanatory power than those for diagnosed groups, suggesting that treatment effects and disease management introduce additional variability not fully captured by observed covariates.

Importantly, the distinction between confidence and prediction intervals underscores a key clinical insight: while mean risk estimates can be determined with high precision in large datasets, individual-level risk remains subject to substantial uncertainty. These findings highlight the importance of combining population-level prediction with individualized clinical judgment when assessing diabetes risk and progression.

6 Conclusion

This study examined how demographic, metabolic, lifestyle, and socioeconomic health factors are associated with diabetes risk across multiple disease types and diagnosis statuses using a large observational dataset. By stratifying the data by diabetes type and diagnosis status and applying rigorous variable selection, influence diagnostics, and model validation procedures, this analysis sought to identify the most stable and interpretable predictors of diabetes risk while accounting for heterogeneity across populations.

6.1 Summary of Findings

Across all disease categories, several consistent patterns emerged. Age was universally associated with increased diabetes risk, reflecting the cumulative physiological and metabolic burden over time. Physical activity emerged as one of the most robust and influential protective factors across nearly all datasets, often exhibiting larger and more stable effects than individual metabolic markers. Lipid measures—particularly HDL cholesterol and triglycerides—played a central role in distinguishing risk profiles, especially in pre-diabetic and Type 2 diabetes populations.

The answers to the primary questions of interest are therefore clear. First, diabetes risk is not driven by a single factor but by a combination of demographic characteristics, lifestyle behaviors, and metabolic indicators. Second, the relative importance of these factors varies substantially by diabetes type and diagnosis status. Models fitted to undiagnosed populations tended to exhibit stronger predictive performance and more stable coefficient estimates, while diagnosed populations—particularly Type 1 diabetes—displayed greater heterogeneity, likely due to treatment effects and disease management strategies not captured in the data. Finally, lifestyle variables such as physical activity and diet quality consistently retained explanatory power even after controlling for biochemical measures, highlighting their central role in diabetes risk.

6.2 Causal Interpretation and Generalizability

Because this analysis is based on observational data, the results should be interpreted as associative rather than causal. While the regression models quantify relationships between predictors and diabetes risk, they do not establish that modifying a given factor will directly cause changes in risk. Confounding, reverse causation, and unmeasured variables—such as medication adherence, healthcare access, or genetic predisposition—may influence the observed associations.

Similarly, caution is warranted when extending these findings to the broader population. Although the dataset is large and diverse, it lacks geographic identifiers and detailed clinical context, and certain groups—most notably undiagnosed Type 2 diabetes—appear underrepresented. As a result, the conclusions are most appropriately generalized to populations with similar demographic and clinical characteristics rather than to all adults in the United States. Nonetheless, the consistency of key findings across multiple subgroups suggests that the identified patterns are likely to be broadly relevant.

6.3 Implications for Policy and Practice

Despite these limitations, the results have important implications for public health policy and clinical practice. The strong and consistent association between physical activity and reduced diabetes risk supports continued investment in programs that promote regular exercise, particularly in populations at risk for pre-diabetes or undiagnosed disease. Community-based interventions, workplace wellness initiatives, and urban planning policies that encourage active lifestyles may yield meaningful reductions in diabetes risk at the population level.

The prominence of lipid measures and diet quality further suggests that nutritional education and access to healthy foods remain critical components of diabetes prevention. Policies aimed at reducing food insecurity, improving nutritional labeling, and expanding access to preventive healthcare services may help mitigate risk before disease progression occurs.

Finally, the finding that undiagnosed populations often exhibit highly predictable risk profiles underscores the importance of early screening and risk stratification. Expanding routine screening for individuals with elevated metabolic risk—particularly older adults and those with poor lifestyle indicators—could improve early detection and reduce the long-term burden of diabetes-related complications.

6.4 Limitations and Future Work

Several limitations of this analysis should be acknowledged, and they point naturally toward directions for future research. First, the models considered in this study were restricted to additive linear effects of predictors. Interaction terms between variables—such as age and physical activity, diet and lipid measures, or insulin and glucose—were not explored. While this choice was made to preserve interpretability, stability, and comparability across datasets with highly variable sample sizes, it necessarily limits the ability of the models to capture effect modification. In reality, the impact of many lifestyle and metabolic factors on diabetes risk is likely conditional on other characteristics, and interaction modeling may reveal important subgroup-specific dynamics.

Relatedly, this analysis did not explicitly compare full models containing all candidate predictors to reduced main-effects-only models beyond those selected through variable selection procedures. Although variable selection implicitly favors parsimonious structures, a systematic comparison of full main-effects models, reduced main-effects models, and interaction-enhanced models could provide additional insight into the trade-offs between explanatory power, prediction accuracy, and interpretability.

Second, the reliance on linear regression imposes structural assumptions that may not fully reflect the underlying biological processes governing diabetes risk. While diagnostic checks indicated acceptable model performance after influence filtering, nonlinear relationships—particularly for glucose, insulin, and lipid measures—are plausible and may be better captured by spline-based or generalized additive models. Future work could explore these alternatives while retaining interpretability.

Third, the observational nature of the dataset precludes causal inference. Although strong and consistent associations were identified, unmeasured confounding and reverse causation remain possible. Incorporating longitudinal data or quasi-experimental designs would allow for stronger causal claims and improved understanding of disease progression over time.

Fourth, several datasets—most notably undiagnosed Type 1 diabetes and gestational diabetes subgroups—contained relatively small sample sizes. This constrained model complexity and limited the feasibility of exploring richer model structures in these populations. Future studies with larger or more targeted samples could allow for more flexible modeling and improved inference in these clinically important but underrepresented groups.

Finally, the dataset lacks detailed clinical and contextual information, such as medication use, duration of disease, healthcare access, and geographic or environmental factors. Including such variables could improve both explanatory power and policy relevance, particularly for understanding disparities in diagnosis and disease management.

Future research should therefore consider extending this framework to include interaction effects, nonlinear modeling strategies, longitudinal designs, and richer clinical covariates. Such extensions would complement the present analysis by deepening insight into the mechanisms underlying diabetes risk while building on the stable and interpretable associations identified here.

6.5 Concluding Remarks

In summary, this analysis demonstrates that diabetes risk is shaped by a complex interplay of age, lifestyle behaviors, and metabolic health, with patterns that differ meaningfully across disease types and diagnosis statuses. While causal conclusions cannot be drawn, the consistency and interpretability of the results provide strong evidence that modifiable lifestyle factors play a central role in diabetes risk across the disease spectrum. These findings reinforce the value of prevention-focused policies and individualized risk assessment in addressing one of the most pressing public health challenges in the United States.

7 Bibliography

References

[1] Centers for Disease Control and Prevention, "Diabetes," 2025. [Online]. Available: https://www.cdc.gov/diabetes/index.html. [Accessed: Dec. 5, 2025].

[2] J. Chatterjee, A. Khunti, and R. Davies, "Type 2 diabetes," *The Lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6482723/. [Accessed: Dec. 5, 2025].

[3] American Heart Association, "What your cholesterol levels mean," 2025. [Online]. Available: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/what-your-cholesterol-levels-mean. [Accessed: Dec. 5, 2025].

[4] M. Krishnathalla, "Diabetes Health Indicators Dataset," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset/discussion/611611. [Accessed: Dec. 5, 2025].

[5] S. W. Lee *et al.*, "Association between changes in systolic blood pressure and incident diabetes: A nationwide population-based study," *Hypertension Research*, vol. 40, no. 8, pp. 710–716, 2017. [Online]. Available: https://www.nature.com/articles/hr201721. [Accessed: Dec. 5, 2025].

[6] S. K. Lee *et al.*, "Lifestyle and metabolic risk factors for diabetes: A comprehensive review," *Journal of Clinical Medicine*, vol. 12, no. 2, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9870675/. [Accessed: Dec. 5, 2025].

[7] M. O'Neill and A. O'Driscoll, "Metabolic syndrome and its relationship with cardiovascular disease and diabetes," *Journal of the American Heart Association*, vol. 8, no. 15, 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6640888/. [Accessed: Dec. 5, 2025].

Table 2: Diabetes Health Indicators Dataset Variables - Numerical and Response

Name	Type	Description	Values/Range	Units/Notes
Demographics				
‘age‘	Integer	Age	18–90	years
Lifestyle				
‘alcohol_use‘	Integer	Weekly alcohol consumption	0–30	drinks per week
‘physical_activity‘	Integer	Weekly physical activity	0–600	minutes per week
‘diet_score‘	Integer	Diet quality	0–10	higher = healthier
‘sleep‘	Float	Average daily sleep	3–12	hours per day
‘screen_time‘	Float	Daily screen time	0–12	hours per day
Physical Health				
‘bmi‘	Float	Body Mass Index	15–45	kilograms per meter squared
‘waist_hip‘	Float	Waist-to-hip ratio	0.7–1.2	abdominal fat indicator
‘systolic_bp‘	Integer	Systolic blood pressure	90–180	millimeters of mercury
‘diastolic_bp‘	Integer	Diastolic blood pressure	60–120	millimeters of mercury
‘hr‘	Integer	Resting heart rate	50–120	beats per minute
Blood Levels				
‘chol_total‘	Float	Total cholesterol	120–300	milligrams per deciliter; Total = LDL + HDL + (Triglycerides / 5)
‘hdl‘	Float	HDL cholesterol	20–100	milligrams per deciliter; High-Density Lipoprotein; good cholesterol
‘ldl‘	Float	LDL cholesterol	50–200	milligrams per deciliter; Low-Density Lipoprotein; bad cholesterol
‘triglycerides‘	Float	Triglycerides	50–500	milligrams per deciliter
‘glucose_fasting‘	Float	Fasting glucose	70–250	milligrams per deciliter
‘glucose_postprandial‘	Float	Post-meal glucose	90–350	milligrams per deciliter
‘insulin‘	Float	Insulin level	2–50	microunits per milliliter
‘hba1c‘	Float	HbA1c	4–14	percent
Diabetic Status				
‘risk‘	Integer	Diabetes risk score	0–100	higher = increased risk, see Table 3 for more details
‘type‘	Factor	Type of diabetes	’No Diabetes’, ’Pre-Diabetes’, ’Type 1’, ’Type 2’, ’Gestational’	
‘diagnosis‘	Factor	Diagnosis indicator	0 = No, 1 = Yes	

Table 3: Diabetes Risk Score Interpretation

Risk Level	Score Range	Interpretation
Low Risk	0–6	Minimal likelihood of diabetes
Moderate Risk	7–11	Moderate chance, lifestyle review recommended
High Risk	12–15	High likelihood, medical screening advised
Very High Risk	16 and above	Very high probability, medical consultation strongly advised

Table 4: Summary of collinearity diagnostics across datasets

Dataset	**Max VIF**	**High Corr. Pairs**	**Cond. Index**	**Severe**	**Notes**
Gestational Diagnosed	< 4	HDL–LDL; TG–Insulin	Moderate	No	Expected metabolic correlations
Gestational Undiagnosed	< 4	HDL–LDL	Moderate	No	Correlations weaker than diagnosed group
No Diabetes Undiagnosed	< 3.5	Minimal	Low	No	Largely independent predictors
Pre-Diabetes Undiagnosed	< 4	TG–Insulin; HDL–LDL	Moderate	No	Metabolic clustering emerging
Type 1 Diagnosed	< 3.5	Minimal	Low	No	Smaller predictor set retained
Type 1 Undiagnosed	< 4	TG–Insulin	Moderate	No	Behavioral and metabolic overlap
Type 2 Diagnosed	< 5	HDL–LDL; TG–Insulin	Moderate	No	Strong metabolic structure; no instability

Table 5: Summary of variable selection results across datasets

Dataset	**Consistently Selected Variables**	**Occasionally Selected Variables**
Gestational Diagnosed	age, physical_activity, HDL, triglycerides, glucose_postprandial	alcohol_use, insulin
Gestational Undiagnosed	age, BMI, physical_activity, HDL, triglycerides, diet_score	screen_time
No Diabetes Undiagnosed	age, diet_score, physical_activity, HDL, triglycerides, insulin, LDL	glucose_postprandial, screen_time
Pre-Diabetes Undiagnosed	age, diet_score, physical_activity, HDL, triglycerides, insulin, LDL	screen_time
Type 1 Diagnosed	age, physical_activity, insulin, diet_score	alcohol_use
Type 1 Undiagnosed	age, physical_activity, insulin, screen_time, triglycerides	diet_score
Type 2 Diagnosed	age, physical_activity, HDL, triglycerides, diet_score, glucose_postprandial, insulin, LDL	alcohol_use, screen_time, sleep

Table 6: Summary of Influential Observation Diagnostics Across Datasets and Models

Dataset	Model	High Leverage	Cook's D	Studentized	Bonferroni
Gestational Diagnosed	Backward p	106	10	5	1
	Forward p	0	2	5	1
	Subsets BIC	106	9	2	1
	Subsets AdjR^2	79	8	7	1
	Subsets RMSE	0	4	5	1
	Subsets C_p	216	7	5	1
Gestational Undiagnosed	Backward p	0	9	11	1
	Forward p	0	8	8	1
	Subsets BIC	0	9	11	1
	Subsets AdjR^2	0	10	11	1
	Subsets RMSE	0	11	11	1
No Diabetes Undiagnosed	Backward p	37,420	388	365	10
	Forward p	0	395	395	1
	Subsets BIC	61,775	386	365	10
	Subsets RMSE	12,191	385	365	10
Pre-Diabetes Undiagnosed	Backward p	1,835,189	3,050	4,092	1
	Forward p	0	1,720	1,720	1
	Subsets AdjR^2	1,390,197	3,119	4,100	1
	Subsets RMSE	778,782	3,185	4,102	1
Type 1 Diagnosed	Backward p	2	3	1	1
	Forward p	0	4	4	1
	Subsets BIC	2	3	1	1
	Subsets RMSE	0	3	1	1
Type 1 Undiagnosed	Backward p	0	5	5	1
	Forward p	0	5	6	1
	Subsets BIC	0	4	4	1
	Subsets RMSE	0	5	5	1
Type 2 Diagnosed	Backward p	3,210,521	1,778	454	1
	Forward p	0	2,323	2,323	1
	Subsets BIC	5,227,472	1,865	460	1
	Subsets AdjR^2	2,484,065	1,721	455	1

Table 7: Validation of Cook's-Distance-Filtered Models

Dataset	Method	n_{orig}	n_{clean}	Removed	% Removed	p	Unstable	Reason
Gestational Undiagnosed	Subsets RMSE	120	51	69	57.50	13	TRUE	$n < 10p$
Gestational Undiagnosed	Subsets AdjR^2	120	69	51	42.50	8	TRUE	$n < 10p$
Gestational Diagnosed	Subsets RMSE	158	103	55	34.81	13	TRUE	$n < 10p$
Type 1 Undiagnosed	Subsets RMSE	56	41	15	26.79	11	TRUE	$n < 10p$
Type 1 Undiagnosed	Backward p	56	44	12	21.43	6	TRUE	$n < 10p$
Type 1 Undiagnosed	Subsets BIC	56	44	12	21.43	5	TRUE	$n < 10p$
Type 1 Diagnosed	Subsets RMSE	66	63	3	4.55	12	TRUE	$n < 10p$
Type 2 Diagnosed	Subsets AdjR^2	59774	36645	23129	38.69	13	FALSE	OK
Type 2 Diagnosed	Subsets BIC	59774	36756	23018	38.51	10	FALSE	OK
Type 2 Diagnosed	Backward p	59774	36769	23005	38.49	12	FALSE	OK
Type 2 Diagnosed	Forward p	59774	45106	14668	24.54	1	FALSE	OK
Pre-Diabetes Undiagnosed	Subsets AdjR^2	31845	23347	8498	26.69	10	FALSE	OK
Pre-Diabetes Undiagnosed	Subsets RMSE	31845	23377	8468	26.59	13	FALSE	OK
Pre-Diabetes Undiagnosed	Backward p	31845	23402	8443	26.51	9	FALSE	OK
Pre-Diabetes Undiagnosed	Forward p	31845	25324	6521	20.48	1	FALSE	OK
No Diabetes Undiagnosed	Backward p	7981	6509	1472	18.44	10	FALSE	OK
No Diabetes Undiagnosed	Subsets RMSE	7981	6529	1452	18.19	13	FALSE	OK
No Diabetes Undiagnosed	Subsets BIC	7981	6545	1436	17.99	9	FALSE	OK
No Diabetes Undiagnosed	Forward p	7981	6690	1291	16.18	1	FALSE	OK
Gestational Diagnosed	Forward p	158	104	54	34.18	1	FALSE	OK
Gestational Diagnosed	Backward p	158	134	24	15.19	6	FALSE	OK
Gestational Diagnosed	Subsets AdjR^2	158	135	23	14.56	8	FALSE	OK
Gestational Diagnosed	Subsets BIC	158	138	20	12.66	4	FALSE	OK
Gestational Diagnosed	Subsets C_p	158	133	25	15.82	5	FALSE	OK
Type 1 Diagnosed	Forward p	66	61	5	7.58	1	FALSE	OK
Type 1 Diagnosed	Backward p	66	63	3	4.55	6	FALSE	OK
Type 1 Diagnosed	Subsets BIC	66	63	3	4.55	3	FALSE	OK

Table 8: Post–Cook's Distance Influence Diagnostics

Dataset	Method	High Leverage	Large Cook's D	Large r-student	Bonferroni
Gestational Diagnosed	Backward p	1	0	12	1
Gestational Diagnosed	Forward p	0	0	1	1
Gestational Diagnosed	Subsets BIC	4	0	8	1
Gestational Diagnosed	Subsets AdjR^2	0	0	12	1
Gestational Diagnosed	Subsets RMSE	0	0	2	1
Gestational Diagnosed	Subsets C_p	1	8	5	1
Gestational Undiagnosed	Backward p	1	0	4	1
Gestational Undiagnosed	Forward p	0	0	1	1
Gestational Undiagnosed	Subsets BIC	0	0	2	1
Gestational Undiagnosed	Subsets AdjR^2	0	0	2	1
Gestational Undiagnosed	Subsets RMSE	0	0	2	1
No Diabetes Undiagnosed	Backward p	1	0	193	1
No Diabetes Undiagnosed	Forward p	0	0	0	1
No Diabetes Undiagnosed	Subsets BIC	4	0	196	1
No Diabetes Undiagnosed	Subsets RMSE	1	0	189	1
Pre-Diabetes Undiagnosed	Backward p	14	0	689	1
Pre-Diabetes Undiagnosed	Forward p	0	0	0	1
Pre-Diabetes Undiagnosed	Subsets AdjR^2	4	0	684	1
Pre-Diabetes Undiagnosed	Subsets RMSE	5	0	626	1
Type 1 Diagnosed	Backward p	0	0	1	1
Type 1 Diagnosed	Forward p	0	0	1	1
Type 1 Diagnosed	Subsets BIC	1	0	1	1
Type 1 Diagnosed	Subsets RMSE	0	0	0	1
Type 1 Undiagnosed	Backward p	0	0	1	1
Type 1 Undiagnosed	Forward p	0	0	2	1
Type 1 Undiagnosed	Subsets BIC	0	0	2	1
Type 1 Undiagnosed	Subsets RMSE	0	0	1	1
Type 2 Diagnosed	Backward p	9	0	1047	1
Type 2 Diagnosed	Forward p	0	0	0	1
Type 2 Diagnosed	Subsets BIC	19	0	1099	1
Type 2 Diagnosed	Subsets AdjR^2	7	0	1023	1

Table 9: Summary of Error Assumption Diagnostics After Influence Filtering

Dataset	Homoscedasticity	Normality	Independence	Notes
Gestational Diagnosed	Mostly Satisfied	Marginal	Satisfied	Mild heteroscedasticity for select metabolic predictors
Gestational Undiagnosed	Mostly Satisfied	Acceptable	Satisfied	Improved substantially after filtering
No Diabetes Undiagnosed	Satisfied	Not Tested[†]	Satisfied	Large sample size; stable residual structure
Pre-Diabetes Undiagnosed	Mostly Satisfied	Not Tested[†]	Satisfied	Minor variance differences for HDL and triglycerides
Type 1 Diagnosed	Mostly Satisfied	Marginal	Satisfied	Small-sample variability; no serial dependence
Type 1 Undiagnosed	Mostly Satisfied	Acceptable	Satisfied	Filtering reduced variance instability
Type 2 Diagnosed	Mostly Satisfied	Not Tested[†]	Satisfied	Minor predictor-specific variance effects

[†] Normality tests were not emphasized for very large samples, where trivial deviations from normality are expected to be statistically significant.

Table 10: Summary of Directional Effects for Key Predictors in Final Models

Predictor	Age	Physical Activity	HDL	Diet Score
Gestational Diagnosed	+	−	−	·
Gestational Undiagnosed	+	−	−	−
No Diabetes Undiagnosed	+	−	−	−
Pre-Diabetes Undiagnosed	+	−	−	−
Type 1 Diagnosed	+	−	·	−
Type 1 Undiagnosed	+	−	−	−
Type 2 Diagnosed	+	−	−	−

+ positive association; − negative association; · not consistently retained

Table 11: Representative Predicted Risk Values with Confidence and Prediction Intervals by Dataset

Dataset	Predicted Risk	95% CI	95% PI
Gestational Diagnosed	≈ 23.3	(23.1, 23.5)	(21.1, 25.5)
Gestational Undiagnosed	≈ 21.8	(21.8, 21.8)	(21.8, 21.9)
No Diabetes Undiagnosed	≈ 23.6	(23.5, 23.6)	(21.5, 25.7)
Pre-Diabetes Undiagnosed	≈ 26.0	(26.0, 26.1)	(24.0, 28.1)
Type 1 Diagnosed	≈ 26.1	(24.3, 27.9)	(11.3, 40.9)
Type 1 Undiagnosed	≈ 19.5	(19.3, 19.9)	(17.5, 21.7)
Type 2 Diagnosed	≈ 28.0	(28.0, 28.0)	(25.9, 30.1)