



Multi-dimensional Observation of Subgroups And Influential Characteristics in Diabetes



Linyuan Yu, Eunsu Hwang, Cao Cong Luan Tran

Summary

Type 2 Diabetes (T2D) remains a major public health burden in the U.S., with substantial personal and economic consequences. Traditional models assume uniform risk factors, yet risk drivers vary sharply across age and income groups. Social determinants, mental health, and healthcare access all shape diabetes risk, but are rarely examined together or across subpopulations. Population-level analyses often mask this heterogeneity, leading to one-size-fits-all interventions. Our work uses interpretable machine learning to uncover subgroup-specific risk patterns and support more targeted, equitable prevention strategies.

Data

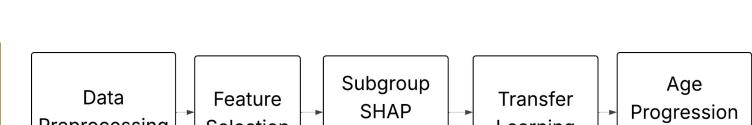
We use the **BRFSS 2024 (Behavioral Risk Factor Surveillance System)**, containing **133,529** complete adult records after cleaning. The outcome variable is **Type 2 diabetes diagnosis**, excluding gestational diabetes and prediabetes.

The dataset includes a wide range of variables across multiple domains:

- Demographics:** age, sex, race, income, education
- Clinical:** BMI category, general health
- Behavioral:** physical activity, smoking
- Mental Health:** depression, mental-health days
- Social Determinants:** food insecurity, transportation barriers, financial strain
- Healthcare Access:** insurance, doctor availability, timing of last checkup

This rich, multidimensional dataset enables comprehensive subgroup and risk-factor analyses across the U.S. adult population.

Method



Feature selection combined **Chi-Square tests** with **SHAP importance** to identify the most influential predictors.

- Association strength between categorical variables and diabetes status
- Trained XGBoost model and computed mean absolute SHAP values

We then trained three complementary machine-learning models to capture diabetes risk patterns across the population:

- LASSO Logistic Regression**
- Random Forest (300 trees)**
- XGBoost (primary model)**

To evaluate early prediction, we performed **transfer learning**, training on age group A_t and testing on the next age group A_{t+1} to assess whether present-day features forecast diabetes 5–10 years later.

For each consecutive age pair (A_t, A_{t+1}) :

- Train model on source age group: $f_t \leftarrow \text{Train}(D_{A_t})$
- Evaluate on target age group: $\text{AUC}_{t \rightarrow t+1} = \text{AUC}(f_t, D_{A_{t+1}})$

Forecasting of age-based prevalence used an ensemble of **Prophet-style decomposition**, **ARIMA**, and **spline smoothing**.

Results

Model Performance

Model	AUC Score
XGBoost	0.811
LASSO Logistic Regression	0.808
Ensemble	0.803
Random Forest	0.803

Model	MAE	RMSE	Bias
Cubic Spline	0.013	0.014	+0.0029
Prophet-style	0.013	0.014	+0.0049
Ensemble	0.011	0.015	+0.002
ARIMA (2,1,0)	0.035	0.035	+0.0107

Source Age	Target Age	AUC (95% CI)
18-24	25-34	0.695 (0.672-0.718)
25-34	35-44	0.780 (0.768-0.792)
35-44	45-54	0.781 (0.771-0.791)
45-54	55-64	0.761 (0.753-0.769)
55-64	65+	0.727 (0.721-0.733)
Mean		0.749 (0.743-0.755)

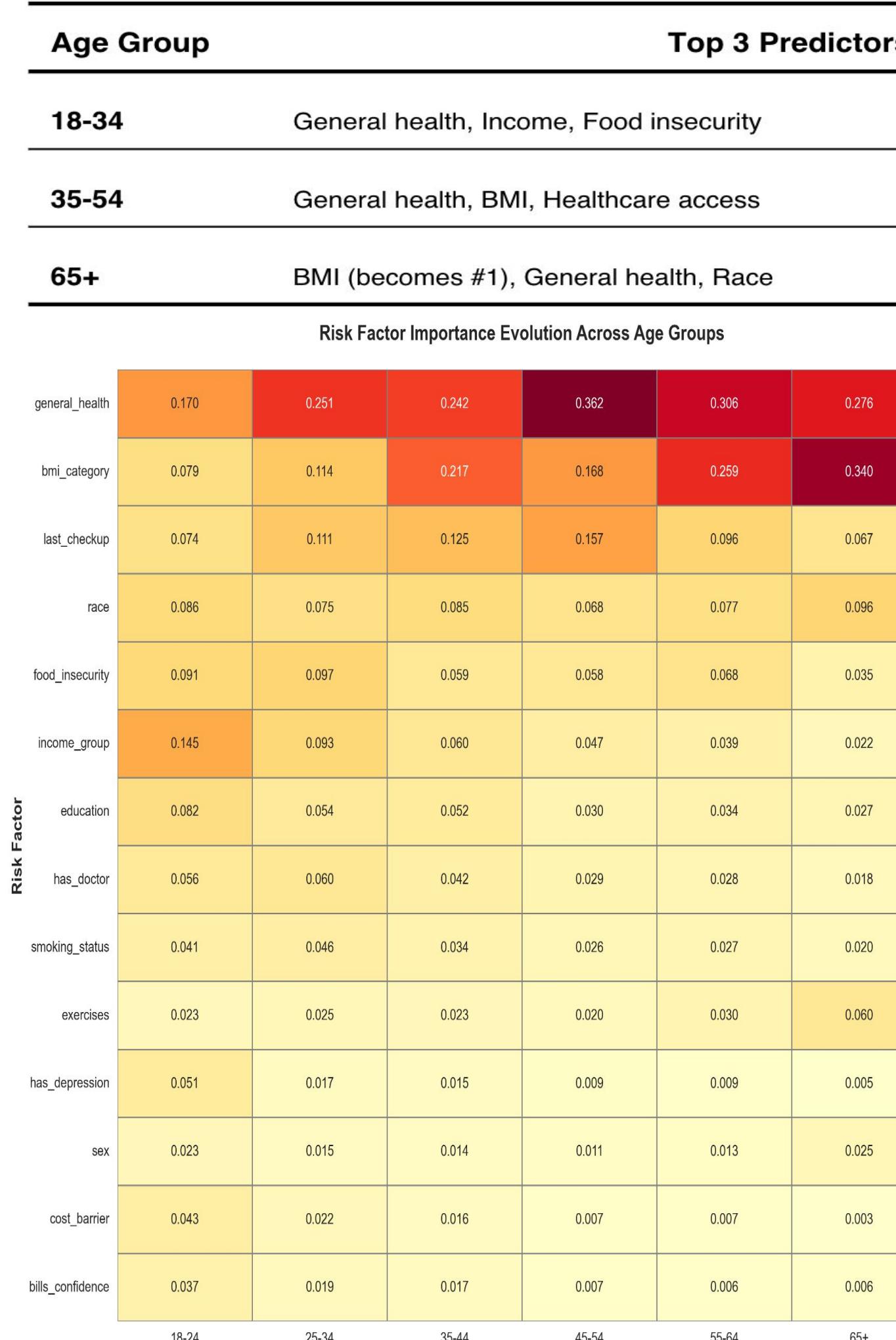
Population Level Feature Importance

Age group, general health, last checkup, income group, and education consistently ranked highly by both methods

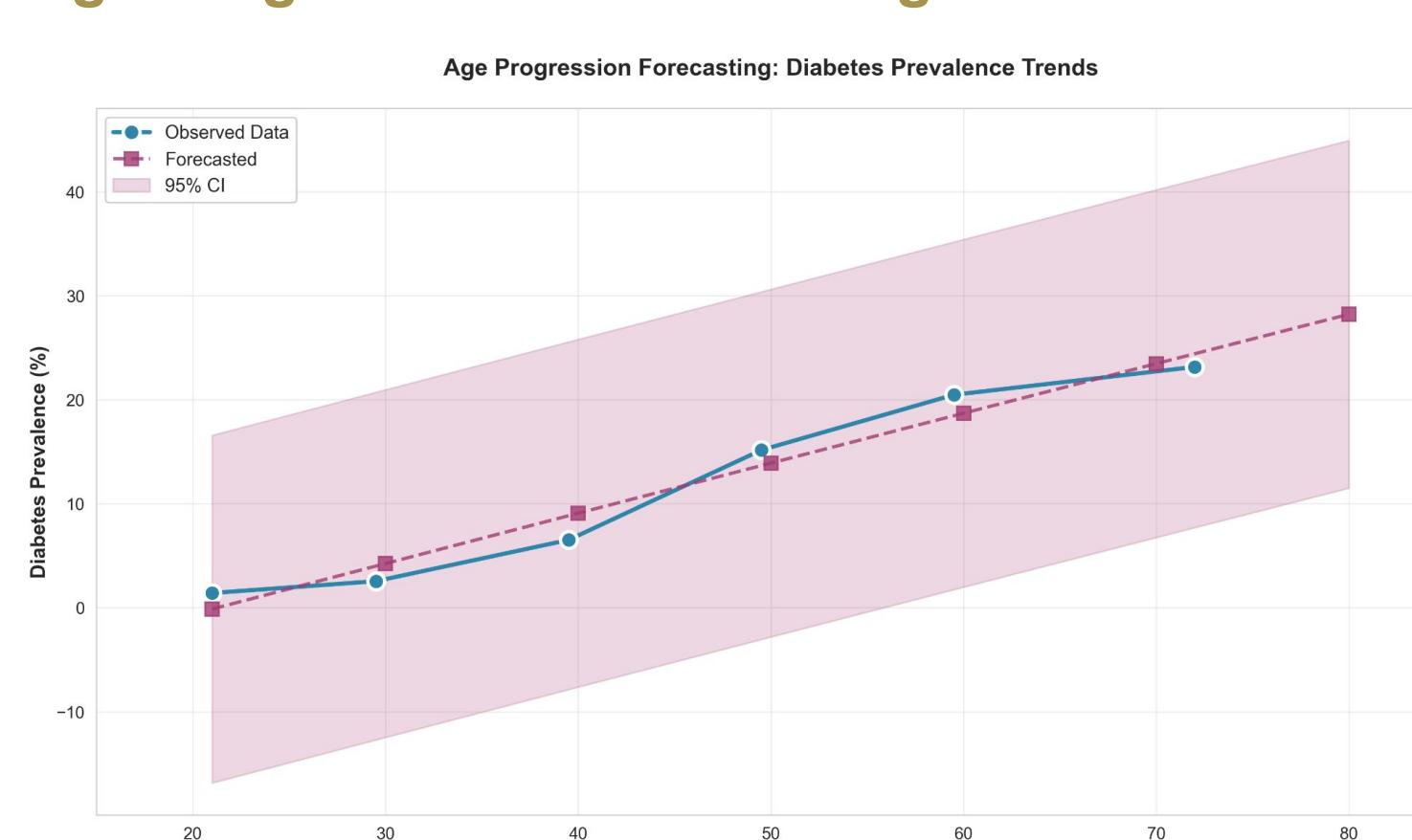
Method	Top 5 features
Chi-squared	Age group, Race, income group, Education, BMI category
XGBoost + Mean SHAP	Age group, General Health, Last checkup, Income, Education

Age-Specific SHAP Analysis

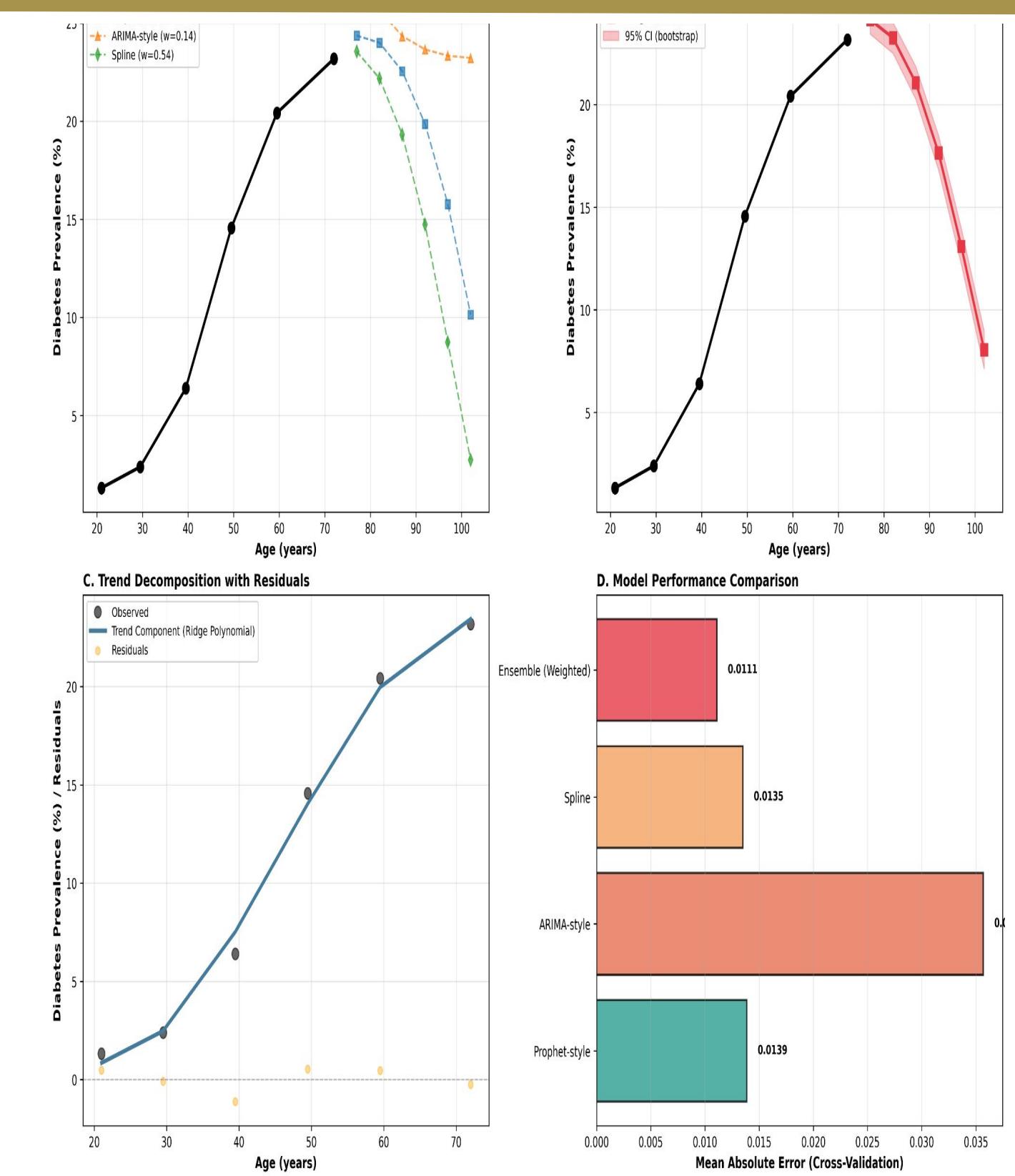
Life-Course Risk Factor Evolution
predictors shift with age — social determinants → BMI → chronic markers



Age Progression + Forecasting

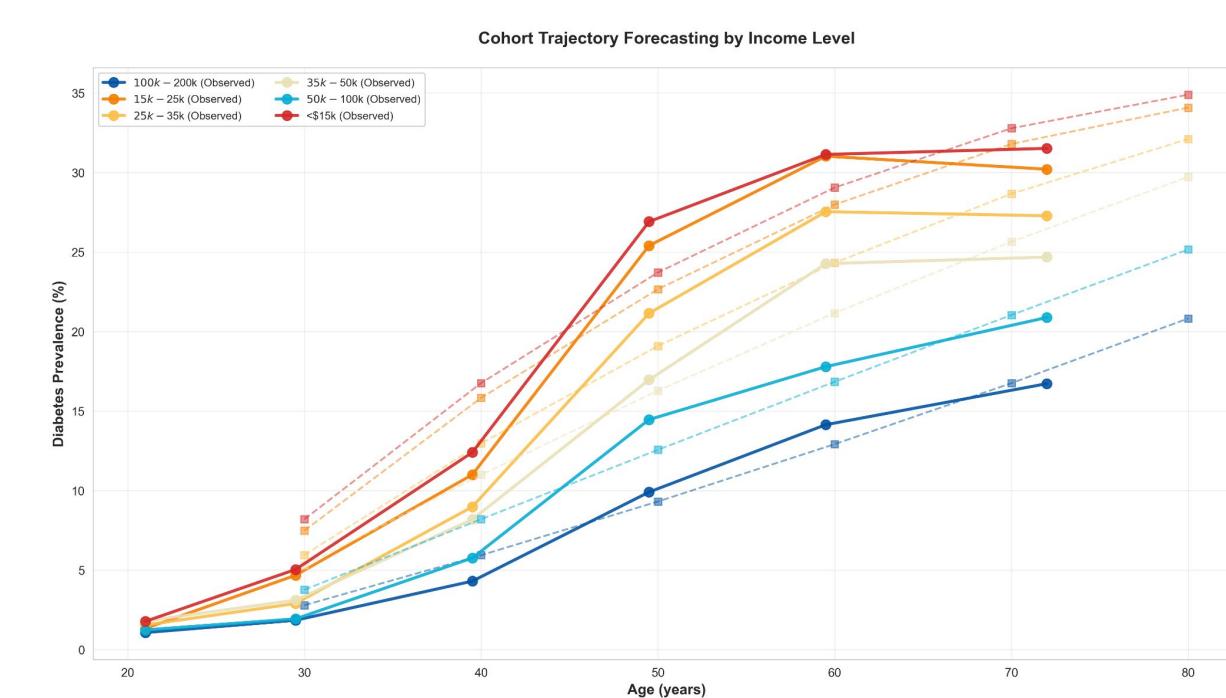


Diabetes prevalence follows a predictable curved pattern as people age, which our mathematical models captured with high accuracy ($R^2 = 0.989$)



Compares three different forecasting methods and their combined "Ensemble" forecast for diabetes rates in people over age 75, showing how uncertainty increases when predicting further into older ages.

Income Trajectories



Diabetes rate gap between the poorest and richest Americans expands 21-fold over the adult lifespan—from 0.7% at age 21 to 14.8% at age 72

Interaction Effects

Predictor	OR	95% CI	p
<i>Main Effects</i>			
Age (per 10 yr)	1.82	1.79-1.85	<0.001
Income (<\$15k)	5.47	4.98-6.01	<0.001
<i>Interaction Terms</i>			
Age × Inc. (<\$15k)	1.23	1.18-1.28	<0.001
Age × Inc. (\$15-25k)	1.19	1.14-1.24	<0.001
Age × Inc. (\$25-35k)	1.15	1.10-1.20	<0.001
Age × Inc. (\$35-50k)	1.11	1.06-1.16	<0.001
Age × Inc. (\$50-100k)	1.07	1.03-1.11	<0.001
Age × Inc. (\$100-200k)	1.04	0.99-1.08	0.089

Note: Reference: Age 18-24, Income ≥\$200k. OR = Odds Ratio. Interaction OR represents multiplicative aging effect within income stratum.

Aging increases diabetes risk more for poor people than for rich people.

Conclusion & Public Health Implications

Diabetes risk changes across adulthood, with its key drivers shifting by age: social determinants in young adults, BMI and healthcare access in midlife, and BMI as the dominant factor in older age. Machine-learning models captured these evolving patterns and accurately predicted risk up to 10 years in advance. Income-stratified analyses showed a dramatic 21-fold widening of risk disparities over adulthood, underscoring the cumulative impact of socioeconomic inequality. These findings support age-specific and income-responsive prevention strategies. Targeting food insecurity and healthcare access during critical windows (ages 25–35 and 40–55) may meaningfully reduce long-term disparities. Integrating interpretable ML with subgroup analysis yields deeper insights and can inform more equitable public health policy.