Data

Source

This report uses data from the **National Health and Nutrition Examination Survey (NHANES)** from 2021-2023. It combines interviews and physical examinations to assess the health and nutritional status of adults and children in the United States.

The merged dataset includes variables representing demographic characteristics, lifestyle factors, and health outcomes. Key variables include:

- **Outcome Variable:** Glycohemoglobin (HbA1c, %), representing blood sugar control and diabetes risk.
- **Main Exposure Variable:** Total daily water intake (grams).
- **Covariates:** Age, gender, race/ethnicity, education level, income-to-poverty ratio (PIR), body mass index (BMI), smoking status, diabetes diagnosis, total energy intake (kcal), and total sugar intake (grams).

The data dictionary

Variable Name

Variable Name (Cleaned)	Original NHANES Code	Description
ID	SEQN	Respondent sequence number
HbA1c	LBXGH	Glycohemoglobin (%) (Outcome / Dependent Variable)
Water_g	DR1TMOIS	Total Daily Moisture Intake (g) (Exposure / Independent Variable)
Age	RIDAGEYR	Age in years at screening
Gender	RIAGENDR	Gender of the respondent
Race	RIDRETH1	Race/Hispanic origin
Education	DMDEDUC2	Education level (Adults 20+)
Income_PIR	INDFMPIR	Ratio of family income to poverty (PIR)
BMI	BMXBMI	Body Mass Index (kg/m**)
Smoking	SMQ020	Smoking history (Smoked at least 100 cigarettes in life)
Diabetes	DIQ010	Doctor diagnosed diabetes
Energy_kcal	DR1TKCAL	Total daily energy intake (kcal)
Sugar_g	DR1TSUGR	Total daily sugar intake (g)

Data dictionary

Categorical data

Variable	Value	Label / Definition
Gender	1	Male
	0	Female
Smoking	1	Yes (Has smoked at least 100 cigarettes in entire life)
	0	No (Has never smoked 100 cigarettes)
Diabetes	1	Yes (Diagnosed with diabetes)
	0	No (Not diagnosed)
Race	1	Mexican American
	2	Other Hispanic
	3	Non-Hispanic White (Commonly used as reference group)
	4	Non-Hispanic Black
	6	Non-Hispanic Asian
	7	Other Race - Including Multi-Racial
Education	1	Less than 9th grade
	2	9-11th grade (Includes 12th grade with no diploma)
	3	High school graduate/GED or equivalent
	4	Some college or AA degree
	5	College graduate or above

Continuous data

Variable	Unit	Notes
HbA1c	%	Clinical cutoff for diabetes is typically ≥ 6.5%
Water_g	Grams (g)	1000 g ≈ 1 Liter
Age	Years	Participants aged 20 and older
BMI	kg/m²	Clinical cutoff for obesity is typically ≥ 30
Income_PIR	Ratio	Range: 0 to 5. Higher values indicate higher socioeconomic status. Values ≥ 1.0 are above the poverty line.
Energy_kcal	kcal	Covariate for total diet quantity
Sugar_g	Grams (g)	Covariate for diet quality

Scraping and Extraction Method

The data is downloaded from the Centers for Disease Control and Prevention's webpage, the National Health and Nutrition Examination Survey. Datasets from different NHANES modules were merged using the respondent ID (SEQN).

Data Cleaning and Processing

First, numerical variables were converted to numeric type to prevent factor-related computation errors. Second, observations with missing values in key variables (HbA1c, Water_g, Energy_kcal, Age, Gender, Race, or BMI) were removed using na.omit(). Third, the categorical variables are recoded:

Gender: 1 = Male, 0 = Female

Smoking: 1 = Smoked ≥100 cigarettes, 0 = Never smoked

Diabetes: 1 = Diagnosed, 0 = Not diagnosed

Race: 1 = Mexican American, 2= Other Hispanic, 3= Non-Hispanic White 4=Non-Hispanic Black

Education: scale 1-5, from "<9th grade" to "College graduate or above."

Water: Q1 corresponds to participants in the **lowest 25%** of daily water intake (lowest hydration group), Q2 and Q3 represent the **middle 50%** of water consumers, and Q4 corresponds to the **highest 25%** of daily water intake (highest hydration group).

Exploratory analysis

This section summarizes the exploratory analyses conducted to understand the relationship between water intake and diabetes outcomes. We examined the distribution of water consumption across diabetes groups, assessed potential confounders, and visualized early associations between hydration and metabolic health. The overall process include discovering overall trend, model examination, sensitivity analysis and stratified visual checks.

Step1: overall trend

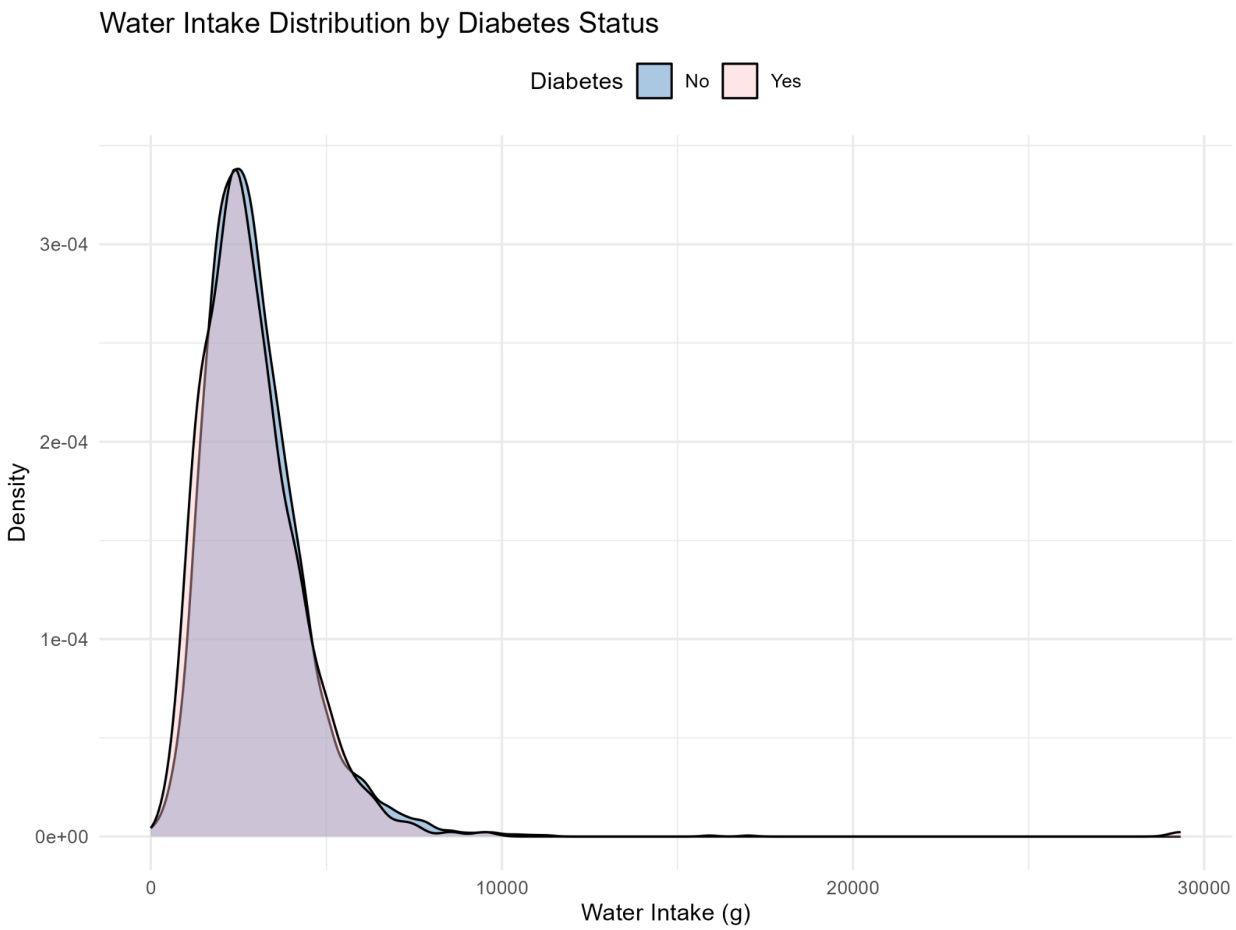

Figure 1

The density plot shows that the overall distribution of daily water intake is highly right-skewed, with most adults consuming between 0–5,000 grams per day. Importantly, the curves for individuals **with** and **without** diabetes nearly overlap. This suggests that **there is no obvious unadjusted difference in water consumption between diabetes groups**.

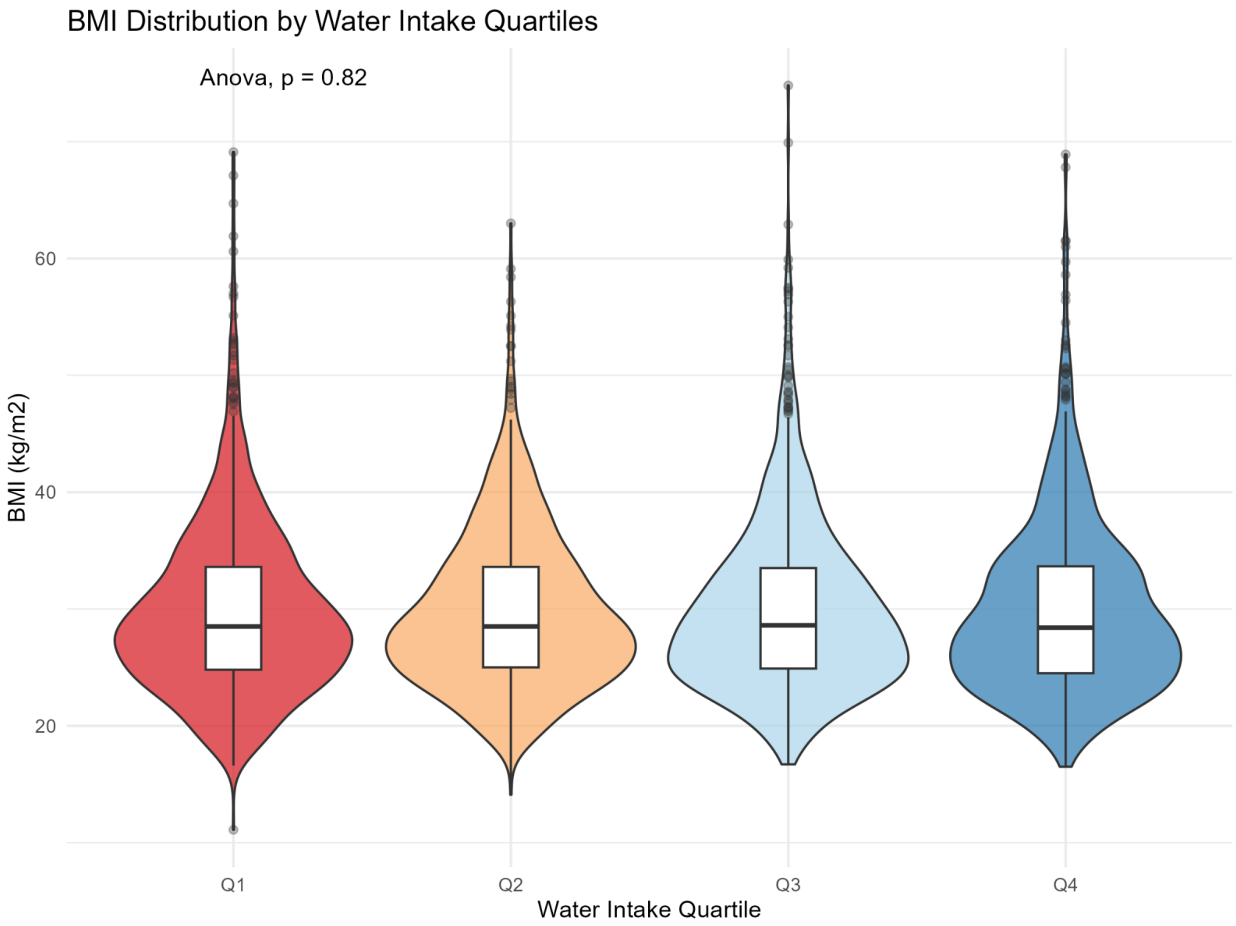

Figure 2

BMI distributions are very similar across the four water-intake quartiles, and the ANOVA test ($p = 0.82$) indicates **no significant differences in BMI** across water consumption levels. This result is important because BMI is a known risk factor for Type 2 diabetes; the absence of systematic BMI differences reduces concern that adiposity confounds the water–diabetes relationship.

This finding also **validated our decision** to include BMI as a covariate but not treat it as an effect modifier at the initial stage.

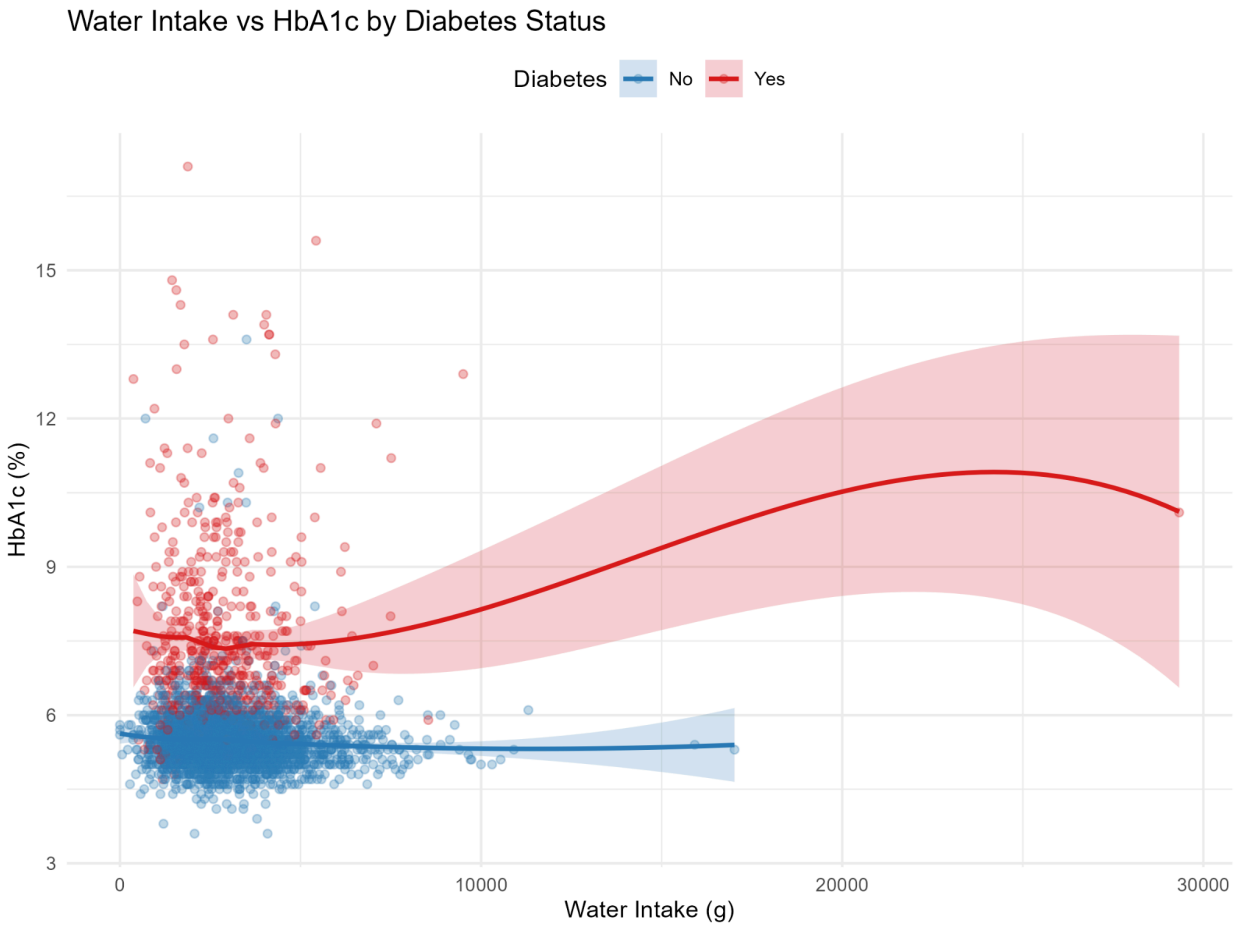

Figure 3

A scatterplot with smoothed curves shows contrasting patterns between diabetes groups.

- For **non-diabetic adults**, HbA1c remains stable across the water-intake range with no apparent trend.
- For **adults with diabetes**, the curve initially slopes downward but rises again at very high consumption values.

The large uncertainty at extreme water intake levels is due to very small sample sizes, which prompted a methodological change: instead of modeling water as a continuous predictor only, we created **water-intake quartiles** to stabilize estimation and facilitate interpretation.

Step2: regression model

To model diabetes status, we fit a sequence of **logistic regression models** with water quartiles as the main exposure:

- **Model 1 (m1, unadjusted):**

$$\text{logit}\{P(\text{Diabetes} = 1)\} = \beta_0 + \beta_1 \text{Water\_Q}.$$

- **Model 2 (m2, demographics-adjusted):**

$$\text{logit}\{P(\text{Diabetes} = 1)\} = \beta_0 + \beta_1 \text{Water\_Q} + \text{Age} + \text{Gender} + \text{Race} + \text{Education} + \text{Income\_PIR}.$$

- **Model 3 (m3, fully adjusted):**

$$\text{logit}\{P(\text{Diabetes} = 1)\} = \beta_0 + \beta_1 \text{Water\_Q} + \beta_2 \text{Age} + \beta_3 \text{Gender} + \beta_4 \text{Race} + \beta_5 \text{Education} + \beta_6 \text{Income\_PIR} + \beta_7 \text{BMI} + \beta_8 \text{Smoking} + \beta_9 \text{Energy\_kcal} + \beta_{10} \text{Sugar\_g}.$$

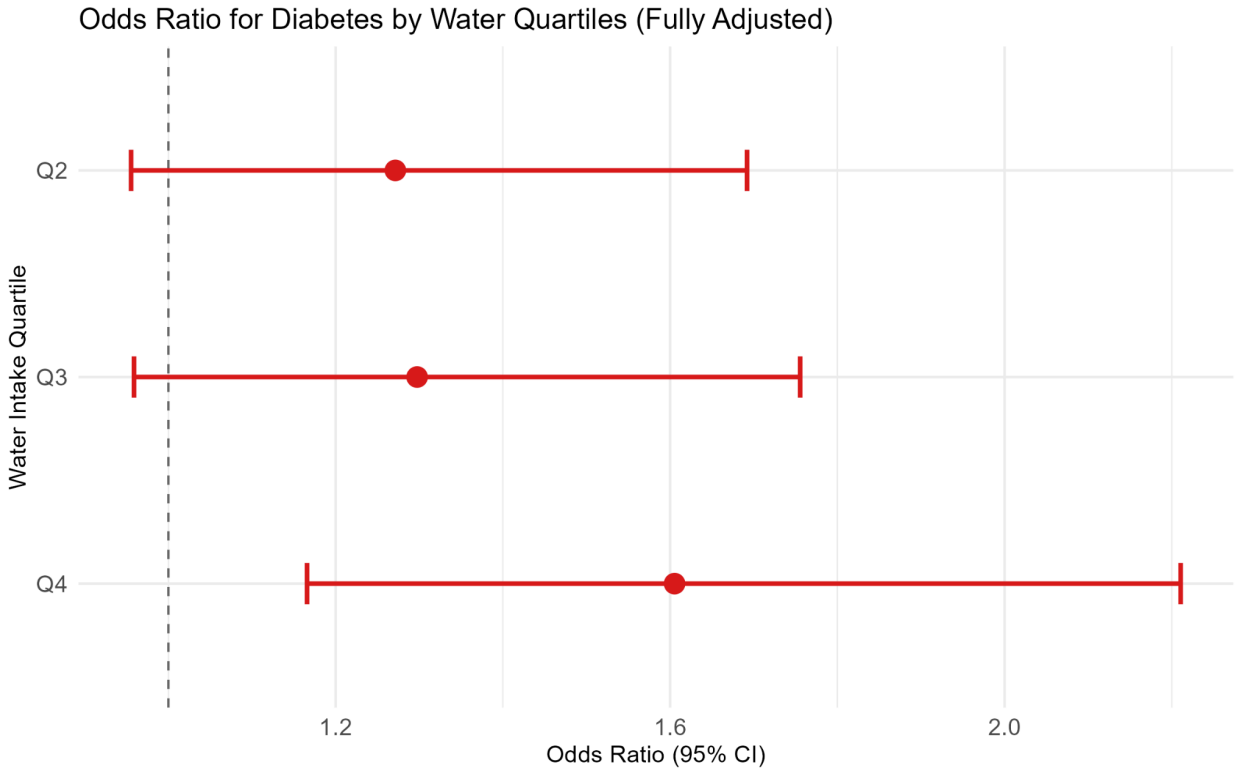

Figure 4

Figure 4 presents the **fully adjusted** odds ratios for Q2–Q4 vs Q1 (reference).

- All odds ratios are **close to 1.0**, with 95% confidence intervals crossing 1.

- The **P for trend** from m3_trend is not significant.

Across the model evolution—unadjusted → demographics-adjusted → fully adjusted—there is **no emergence of a protective association**. If anything, estimates remain near null or slightly above 1, suggesting **no evidence that higher water intake reduces diabetes odds**.

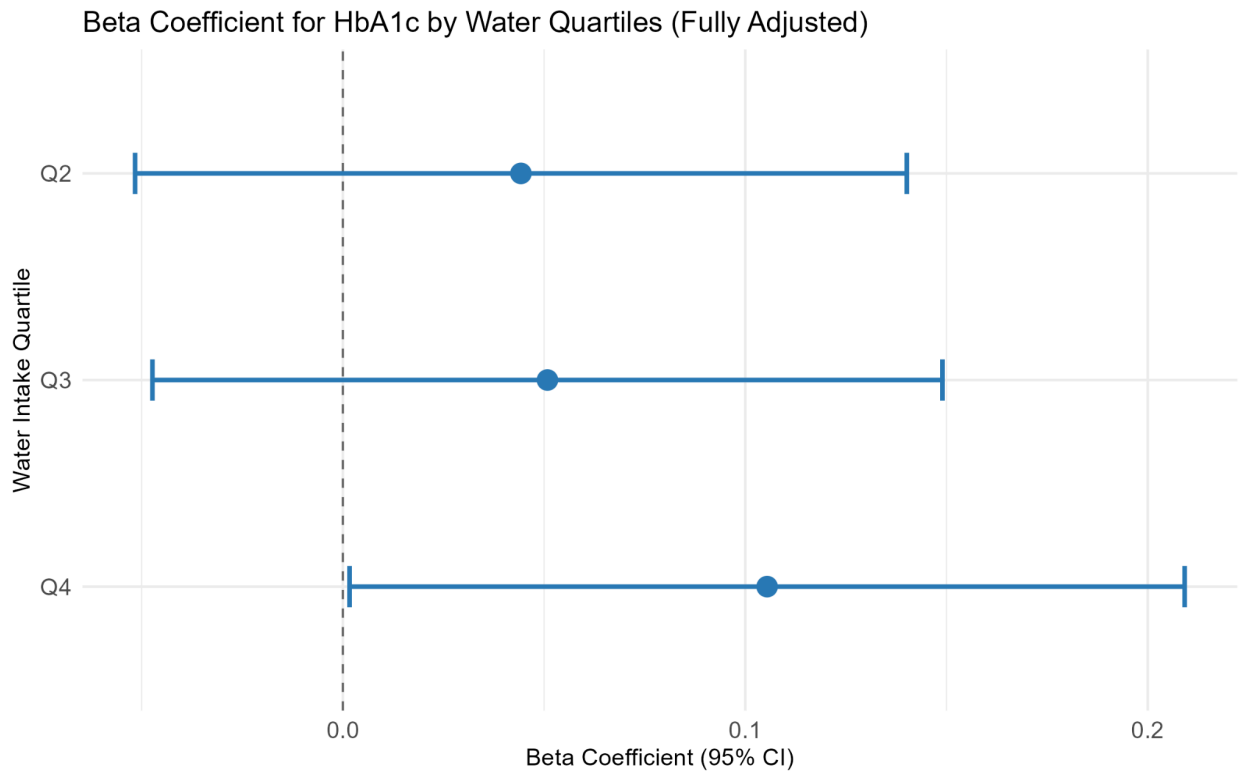

Figure 5

Figure 5 shows the **fully adjusted** beta coefficients for Q2–Q4 vs Q1.

- All estimates are very close to **0**, and confidence intervals include 0.
- The **trend test** is non-significant in the fully adjusted model.

Across unadjusted, demographics-only, and fully adjusted models, we consistently see **no meaningful difference in HbA1c by water-intake quartile**.

Step 3: sensitivity analysis

To check robustness, we reanalyzed water as a **continuous variable**:

- **Logistic sensitivity model (m_cont):**

  $$\text{logit}\{P(\text{Diabetes} = 1)\} = \beta_0 + \beta_1 \text{Water\_g} + \text{covariates (as in m3)}.$$

  The result was summarized as an **OR per 100 g** increase in Water_g.

- **Linear sensitivity model (lm_cont):**

  $$\text{HbA1c} = \alpha_0 + \alpha_1 \text{Water\_g} + \text{covariates (as in lm3)} + \varepsilon,$$

In both models, the **per-100 g effect was extremely small and not statistically significant**, consistent with the quartile-based results and confirming that treating water as continuous does not reveal a hidden dose–response association.

Step 4:descriptive and stratified visual checks

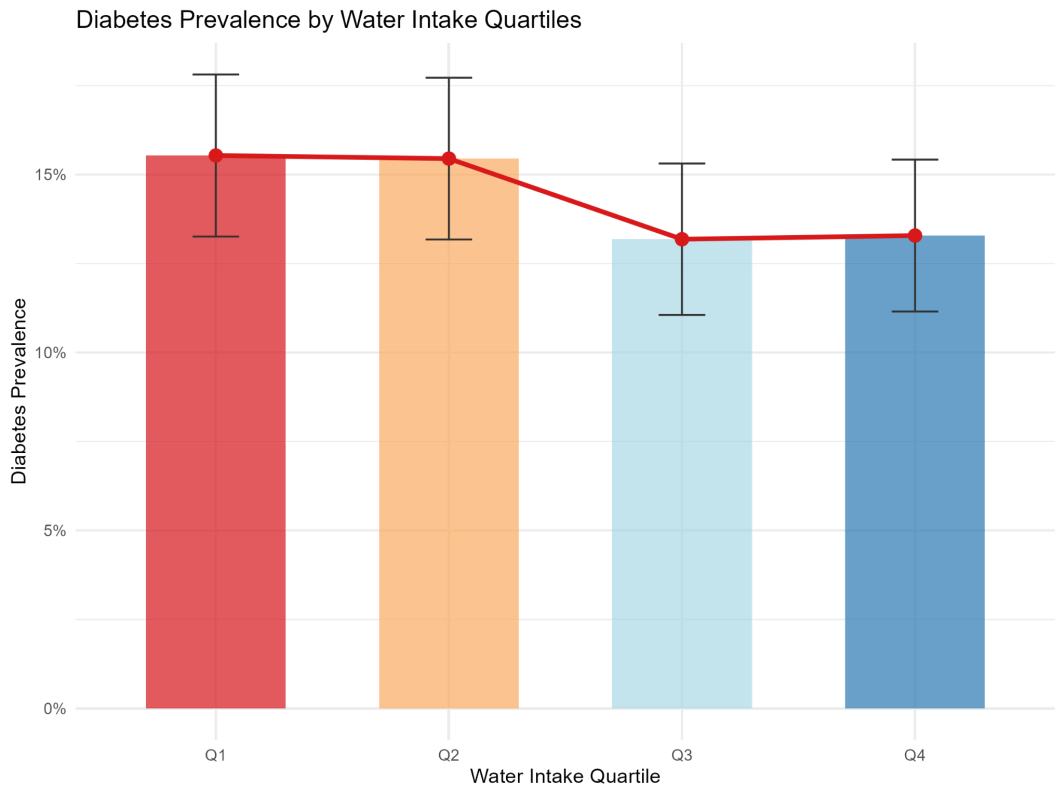

Figure 6

Figure 6 shows the crude diabetes prevalence across Water_Q.

- Although prevalence decreases slightly from Q1 to Q4, the wide and overlapping confidence intervals indicate no meaningful unadjusted association.

- This crude pattern is fully consistent with the **logistic regression models (m1–m3)**, where all quartile odds ratios remain close to 1 and non-significant, confirming no protective effect of higher water intake.

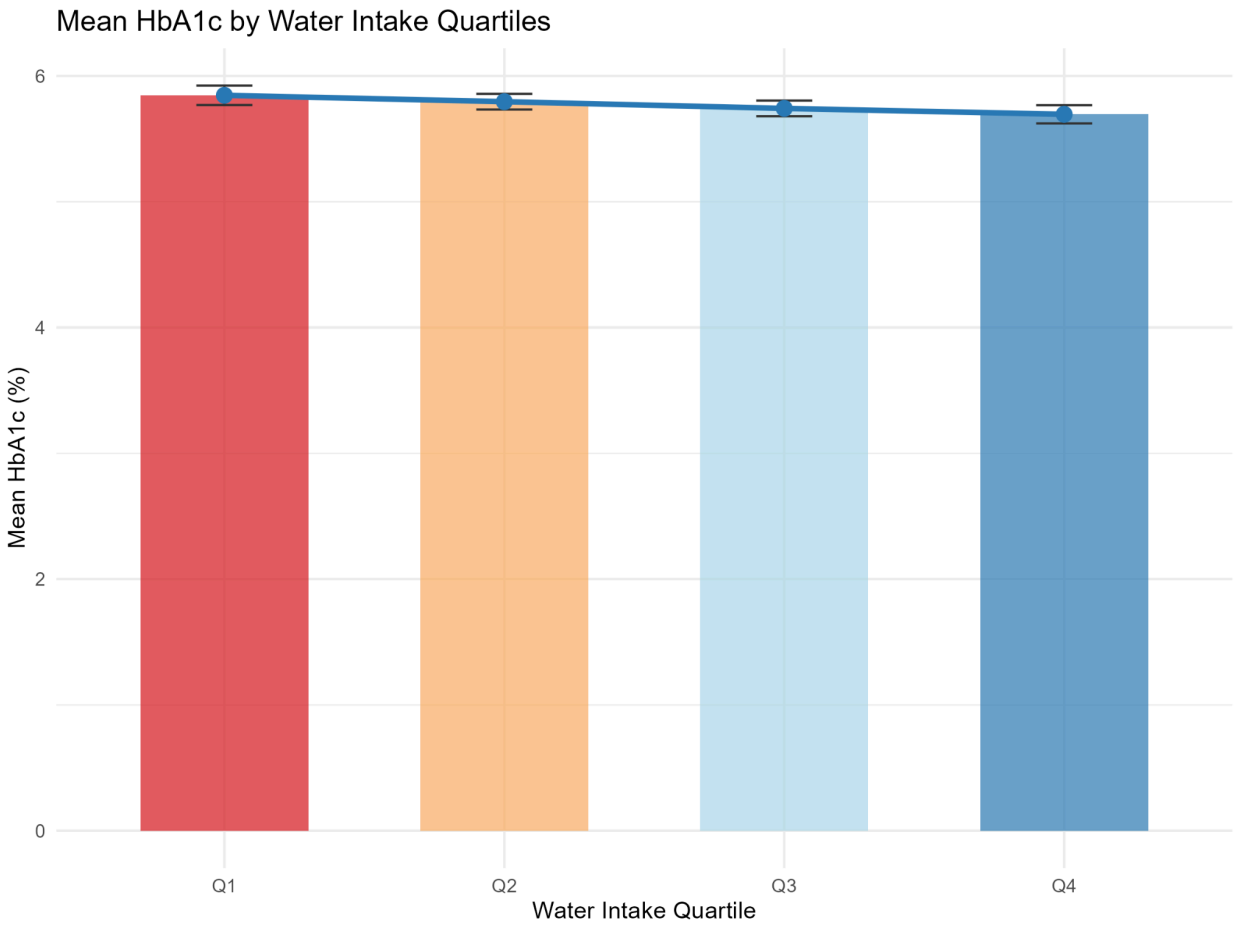

Figure 7

Figure 7 presents mean HbA1c with 95% confidence intervals across quartiles.

- HbA1c declines modestly in higher quartiles, but absolute differences are small and confidence intervals heavily overlap.

- This descriptive pattern aligns with **linear regression models (lm1–lm3)**, where all quartile beta estimates were near zero with non-significant trend tests, indicating that hydration level does not meaningfully predict HbA1c.

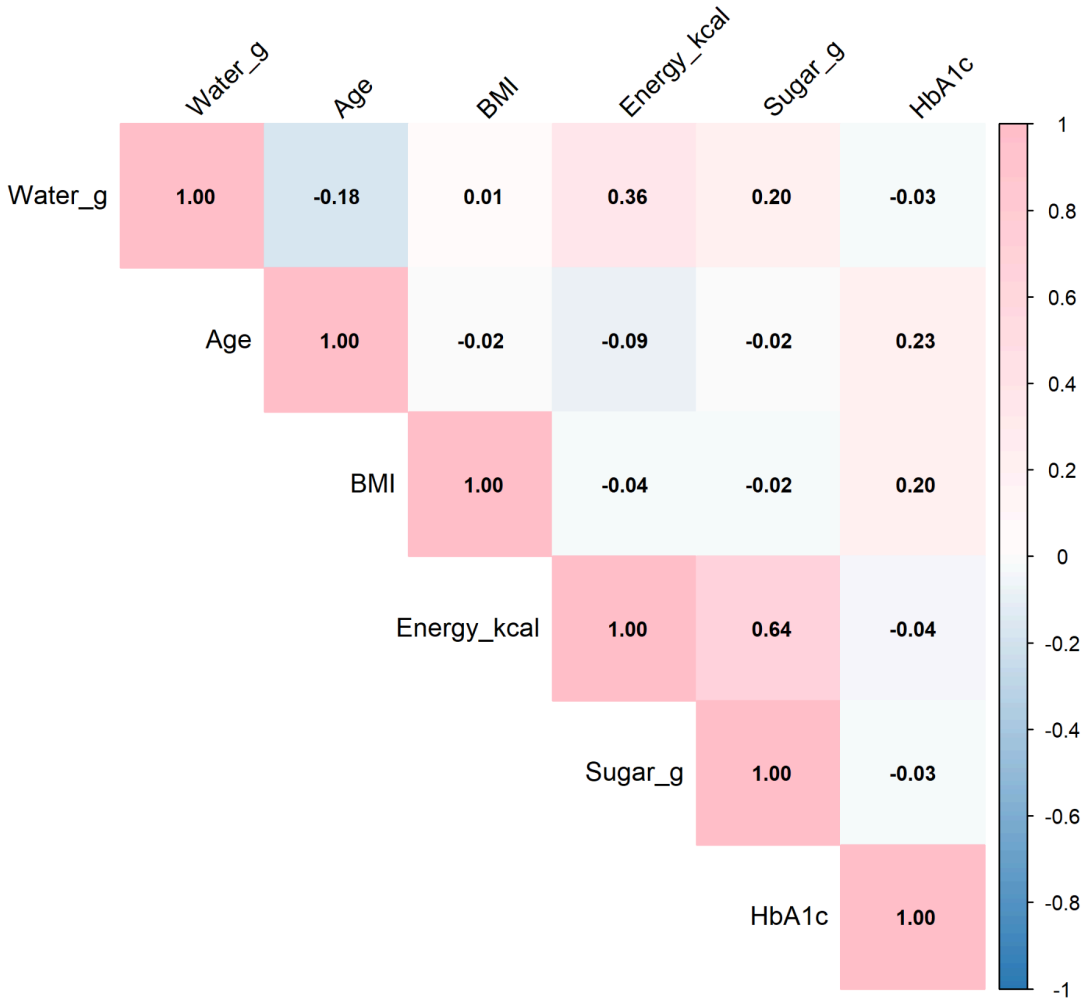

Figure 8

Figure 8 shows Pearson correlations among Water_g, Age, BMI, Energy_kcal, Sugar_g, and HbA1c.

- Correlations between Water_g and other variables, including HbA1c, are **very weak** (absolute values close to 0).

- Energy_kcal and Sugar_g show a moderate positive correlation, as expected from dietary data.

These correlations reassure us that **multicollinearity is low** and that water intake itself contributes little variation to HbA1c or diabetes status.

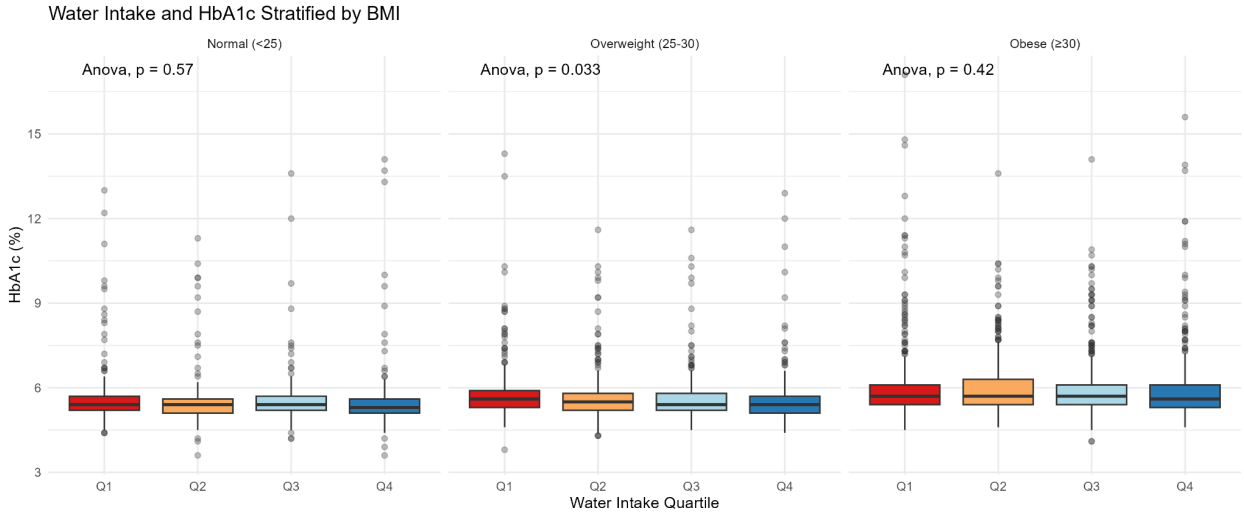

Figure 9

Finally, **Figure 9** explores whether BMI modifies the water–HbA1c relationship. HbA1c is plotted by Water_Q within BMI categories (Normal, Overweight, Obese), and ANOVA p-values are shown in each panel.

- **Normal weight:** no significant differences in HbA1c across water quartiles.

- **Overweight:** a small statistically significant difference ($p \approx 0.03$), with slightly lower HbA1c in higher quartiles.

- **Obese:** no significant differences.