Survey-Weighted Estimation of Diagnosed Diabetes Prevalence and Its Correlates in U.S. Adults

A Design-Based Analysis of NHANES 2017–2018 Using a Manual Taylor-Linearization Sandwich Variance Estimator

July 10, 2026

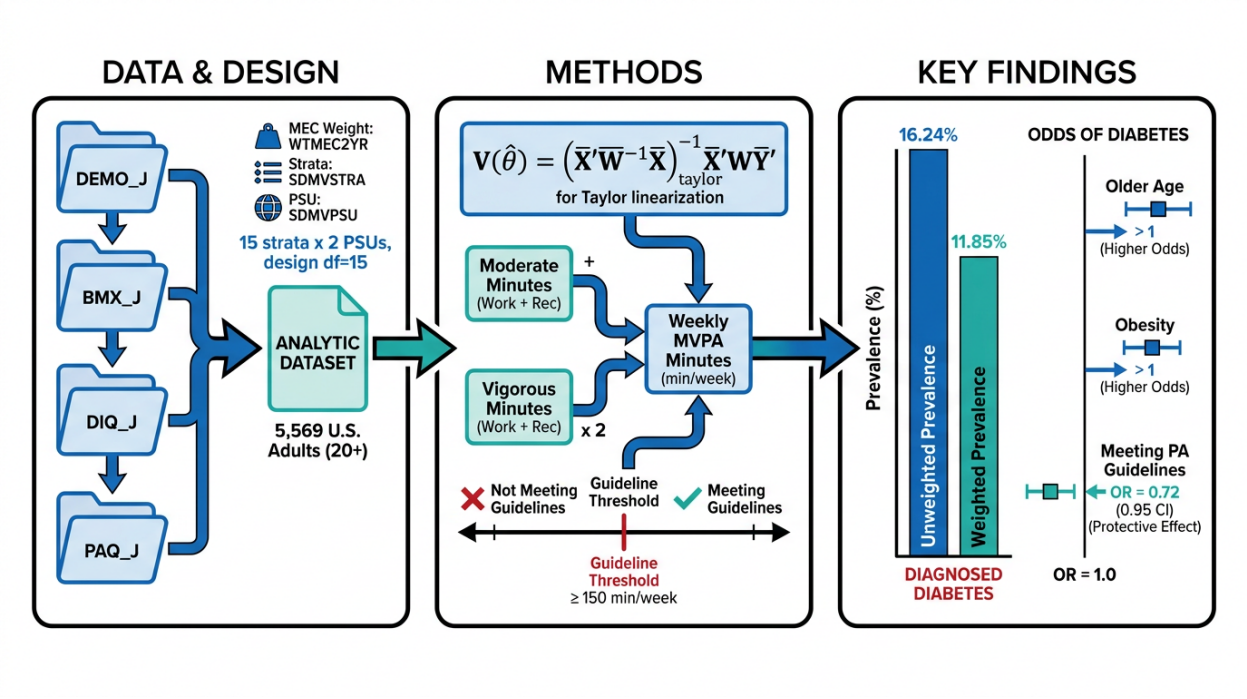

Figure 1: Graphical abstract. The analytic workflow links four NHANES 2017–2018 public-use files by respondent sequence number (`SEQN`), incorporates the complex survey design (MEC weight `WTMEC2YR`, masked strata `SDMVSTRA`, masked PSUs `SDMVPSU`; design df = 15), derives weekly moderate-to-vigorous physical activity (MVPA) with vigorous minutes double-weighted, and produces design-based estimates. Key findings: ignoring the survey design inflates the diagnosed-diabetes prevalence estimate from 11.85% to 16.24%, and the adjusted model shows strong age and adiposity gradients with a protective association for meeting the physical-activity guideline.

Executive Summary

Diagnosed diabetes is one of the most consequential chronic conditions in the United States, and its national surveillance depends almost entirely on the National Health and Nutrition Examination Survey (NHANES). Because NHANES is a stratified, multistage, clustered probability sample that deliberately oversamples older adults and several higher-risk subpopulations, valid national inference *requires* the sampling weights, masked strata, and masked primary sampling units (PSUs) to be used together. This report documents a fully reproducible, design-based analysis of the NHANES 2017–2018 cycle restricted to adults aged $\geq$ 20 years.

Four public-use files were merged on the respondent sequence number `SEQN`: the demographics file `DEMO_J` (design variables `WTMEC2YR`, `SDMVSTRA`, `SDMVPSU`; age `RIDAGEYR`; family income-to-poverty ratio `INDFMPIR`), the body-measures file `BMX_J` (measured BMI `BMXBMI`), the diabetes questionnaire `DIQ_J` (diagnosed diabetes `DIQ010`), and the physical-activity file `PAQ_J`. Because no first-class complex-survey package was available in the computing environment, the Taylor-series linearization ("sandwich") variance estimator used by R's `survey::svymean` and `svyglm` was implemented directly and validated.

Headline findings. The survey-weighted national prevalence of diagnosed diabetes among U.S. adults was **11.85%** (95% CI: 10.70–13.01), closely matching independent CDC national estimates. Ignoring the design produced a naive prevalence of **16.24%** (95% CI: 15.26–17.23)—an upward bias of roughly **4.4 percentage points**—because the unweighted sample overrepresents older and higher-risk adults. In the survey-weighted multivariable logistic regression, the odds of diagnosed diabetes rose steeply with age (60+ vs. 20–39: adjusted odds ratio [aOR] **13.57**) and adiposity (obese vs. normal BMI: aOR **4.43**); meeting the $\geq$150 min/week MVPA guideline was protective (aOR **0.72**); and low family income was associated with higher odds (aOR **1.43**). These results reproduce the established epidemiology of type 2 diabetes and, more importantly, demonstrate concretely why the survey design cannot be ignored.

Contents

1 Introduction

Diabetes mellitus is among the most prevalent, costly, and consequential chronic diseases in the United States and worldwide. Global surveillance places the number of adults living with diabetes at more than half a billion, with the Global Burden of Disease Study estimating 529 million cases in 2021 and projecting continued growth through 2050 [1], while the International Diabetes Federation similarly estimated roughly 537 million affected adults aged 20–79 in 2021 [2, 3]. Within the United States, nationally representative NHANES analyses have repeatedly shown that total diabetes affects on the order of one in seven to one in eight adults, with age-standardized prevalence climbing from about 9.8% in 1988–1994 to roughly 14% by 2017–2018 [4, 5]. The clinical stakes are high: diabetes is a strong, independent risk factor for coronary heart disease, stroke, and vascular death [6], and the disease imposed an estimated $327 billion in direct medical costs and lost productivity in the United States in 2017 alone [7]. Despite decades of therapeutic advances, achievement of simultaneous glycemic, blood-pressure, and lipid targets has plateaued or worsened in recent NHANES cycles [8, 5], underscoring the continued importance of accurate population surveillance to guide policy and resource allocation.

Accurate national estimates of diagnosed diabetes depend not only on high-quality measurement but also on the correct statistical treatment of the survey that produces them. NHANES is not a simple random sample; it is a complex, stratified, multistage, clustered probability sample of the civilian non-institutionalized U.S. population, and by design it oversamples specific subpopulations—older adults, several race and ethnicity groups, and lower-income households—to ensure adequate precision for subgroup estimates [9, 10]. Consequently, the raw (unweighted) composition of the NHANES sample is deliberately *not* representative of the national population. Producing valid national estimates therefore requires three design elements to be used together: the sampling weight, which maps each examined participant back to the number of Americans they represent; the masked pseudo-stratum (`SDMVSTRA`); and the masked pseudo-primary-sampling-unit (`SDMVPSU`), which together encode the clustering needed for correct variance estimation [9, 11, 12].

The consequences of ignoring these features are well documented in the survey-statistics literature. Weighted and unweighted point estimates answer subtly different questions, and the choice of whether and how to weight has real implications for both bias and efficiency [13, 11]. Ignoring the clustering and stratification—treating the data as if they arose from simple random sampling—typically produces standard errors that are too small and confidence intervals that are too narrow, because positively correlated observations within sampled clusters carry less independent information than their nominal count suggests [12, 11]. Design-based inference addresses both problems

simultaneously by weighting the point estimate to the target population and by using a variance estimator, most commonly Taylor-series linearization, that respects the stratified clustered structure of the sample [12].

The present report has two complementary objectives. The first is substantive: to estimate the weighted national prevalence of diagnosed diabetes among U.S. adults in NHANES 2017–2018 with a correct 95% confidence interval, and to characterize its adjusted associations with age, measured body-mass index (BMI), physical-activity guideline attainment, and family income through a survey-weighted multivariable logistic regression. The second is methodological and pedagogical: to quantify, side by side, how much the prevalence estimate and its uncertainty change when the survey design is ignored, and to document a transparent, from-scratch implementation of the Taylor-linearization sandwich variance estimator so that every number in the report is fully reproducible. Physical-activity status is operationalized against the widely endorsed guideline of at least 150 minutes per week of moderate-to-vigorous physical activity [14, 15, 16], with the exact derivation stated in full. Together these analyses reproduce the known epidemiology of type 2 diabetes while providing a concrete, auditable demonstration of why the complex survey design must be honored.

2 Methods

2.1 Data source and files

All data are drawn from the 2017–2018 ("J") cycle of the continuous NHANES, a program of the U.S. National Center for Health Statistics (NCHS). Four public-use SAS transport (`.XPT`) files were downloaded from the NCHS public data repository and merged one-to-one on the respondent sequence number `SEQN`. Table 1 lists the exact files and the variables extracted from each. The body-measures file supplies *measured* BMI from the Mobile Examination Center (MEC); accordingly, the MEC examination weight `WTMEC2YR` is the correct sampling weight for all analyses, because every analytic variable used here (BMI in particular) was collected in the MEC subsample [9].

Table 1: NHANES 2017–2018 public-use files and variables used in the analysis.

File	Content	Variables used	Role in analysis
`DEMO_J`	Demographics & survey design	`SEQN`, `RIDAGEYR`, `INDFMPIR`, `WTMEC2YR`, `SDMVSTRA`, `SDMVPSU`	Age restriction and age group; family-income category; MEC weight; masked stratum; masked PSU
`BMX_J`	Body measures	`SEQN`, `BMXBMI`	Measured BMI → WHO BMI category
`DIQ_J`	Diabetes questionnaire	`SEQN`, `DIQ010`	Diagnosed diabetes (outcome)
`PAQ_J`	Physical activity	`SEQN`, `PAQ605/610`, `PAD615`, `PAQ620/625`, `PAD630`, `PAQ650/655`, `PAD660`, `PAQ665/670`, `PAD675`	Weekly MVPA minutes → guideline status

2.2 Study population

Analyses were restricted to adults aged ≥ 20 years (`RIDAGEYR` ≥ 20), consistent with the age threshold used in standard NHANES diabetes-surveillance analyses [4, 5]. Of 9,254 examined participants in the cycle, 5,569 met the adult age criterion and formed the analytic cohort. Figure 2 summarizes participant flow and the analytic samples used for each objective. Prevalence estimation used the domain of adults with a non-missing diabetes outcome (n = 5,393); the multivariable regression used complete cases across the outcome and all four covariates (n = 4,328). Missingness in `BMXBMI` (7.1%) and `INDFMPIR` (14.2%) is inherent to NHANES and is the primary driver of the reduction to complete cases.

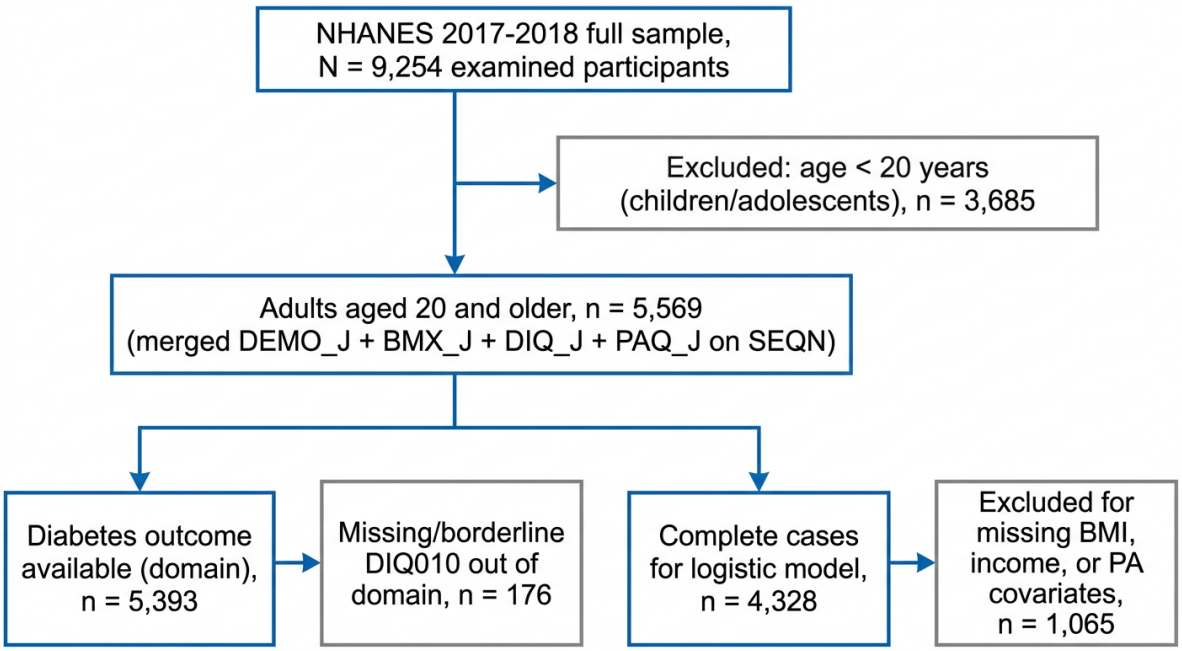

Figure 2: Analytic sample derivation. From the full 2017–2018 examined sample, adults aged ≥ 20 years were retained and the four files merged on `SEQN`. The prevalence analysis uses the outcome domain (n = 5,393); the logistic regression uses complete cases across the outcome and all covariates (n = 4,328). Out-of-domain units are retained (contributing zero) in the linearized variance so that the strata/PSU structure is preserved (domain estimation).

2.3 Outcome and covariate definitions

The outcome, *diagnosed diabetes*, was defined from `DIQ010` ("Doctor told you have diabetes"): a response of "Yes" (code 1) was coded 1 and "No" (code 2) was coded 0. Borderline/prediabetes (code 3), "don't know" (code 9), and refused/missing responses were set to missing and excluded from the outcome domain. This self-reported, physician-diagnosed definition is the standard operationalization of *diagnosed* (as opposed to total or undiagnosed) diabetes in NHANES surveillance [4, 5, 17].

Covariates were operationalized as follows. *Age group* was derived from `RIDAGEYR` as 20–39, 40–59, and 60+ years (reference: 20–39). *BMI category* was derived from measured BMI (`BMXBMI`) using standard World Health Organization cut-points: underweight (< 18.5), normal (18.5 to < 25), overweight (25 to < 30), and obese (≥ 30 kg/m^2) (reference: normal) [18]. *Family-income category* was derived from the family income-to-poverty ratio `INDFMPIR` as low (< 1.3), middle (1.3 to ≤ 3.5), and high (> 3.5) (reference: high). *Physical-activity status* was derived from `PAQ_J`

as meeting versus not meeting the ≥150 min/week MVPA guideline (reference: not meeting), as detailed next.

2.4 Derivation of weekly moderate-to-vigorous physical activity (MVPA)

The Global Physical Activity Questionnaire module in `PAQ_J` captures four activity domains, each with a yes/no screener, a number of days per week, and typical minutes per day: vigorous-intensity work, moderate-intensity work, vigorous-intensity recreation, and moderate-intensity recreation. In the 2017–2018 cycle the *days* items carry the `PAQ` prefix (`PAQ610`, `PAQ625`, `PAQ655`, `PAQ670`) while the *minutes* items carry the `PAD` prefix (`PAD615`, `PAD630`, `PAD660`, `PAD675`); the screeners are `PAQ605`, `PAQ620`, `PAQ650`, and `PAQ665`. For each domain, a screener response of "No" (code 2) set that domain's days and minutes to zero; invalid day codes (77, 99) and invalid minute codes (7777, 9999) were set to missing.

Weekly minutes for each domain were computed as (days per week) × (minutes per day). Total weekly MVPA minutes were then formed by summing the four domains, with *vigorous* minutes double-weighted to reflect their approximately twofold metabolic intensity relative to moderate activity—the standard convention embedded in the physical-activity guidelines, under which 75 minutes of vigorous activity is treated as equivalent to 150 minutes of moderate activity [14, 15]. Formally,

$$\mathrm{MVPA}_{\mathrm{min/wk}} = \underbrace{(d^{\mathrm{WM}} \cdot m^{\mathrm{WM}})}_{\text{work, moderate}} + \underbrace{2\,(d^{\mathrm{WV}} \cdot m^{\mathrm{WV}})}_{\text{work, vigorous}} + \underbrace{(d^{\mathrm{RM}} \cdot m^{\mathrm{RM}})}_{\text{recreation, moderate}} + \underbrace{2\,(d^{\mathrm{RV}} \cdot m^{\mathrm{RV}})}_{\text{recreation, vigorous}}, \tag{1}$$

where $d^{\bullet}$ and $m^{\bullet}$ denote weekly days and daily minutes for each domain, with superscripts W/R (work/recreation) and M/V (moderate/vigorous). Mapping Equation (1) to variable names: work moderate uses (`PAQ625`, `PAD630`), work vigorous uses (`PAQ610`, `PAD615`), recreation moderate uses (`PAQ670`, `PAD675`), and recreation vigorous uses (`PAQ655`, `PAD660`). Respondents were classified as *meeting* the guideline when $\mathrm{MVPA}_{\mathrm{min/wk}} \geq 150$ and *not meeting* otherwise [14, 15, 16]. Respondents who screened positive for a domain but refused or did not know the corresponding days/minutes were left as missing for MVPA (43 adults), a conservative choice.

2.5 Survey design and design-based estimation

The 2017–2018 cycle contains 15 masked pseudo-strata (`SDMVSTRA`), each with two masked pseudo-PSUs (`SDMVPSU`), giving 30 PSUs and hence a design degrees of freedom of df = (number of PSUs) − (number of strata) = 30 − 15 = 15 [9]. All design-based confidence intervals and p-values use a t-distribution with 15 degrees of freedom.

Because no first-class complex-survey package (e.g., R's `survey`, or a Python equivalent) was available in the computing environment, the Taylor-series linearization variance estimator used by `survey::svymean` and `svyglm` was implemented directly [12]. For a stratified, single-stage, with-replacement design—the standard NHANES analytic approximation using the masked design variables—any statistic defined through weighted estimating equations $U = \sum_i w_i\, u_i$ has a linearization variance of the sandwich form

$$\widehat{\mathrm{Var}} \;=\; A^{-1}\, B\, A^{-1}, \tag{2}$$

where for a simple mean or proportion $A = 1$ and the estimator reduces to the "meat" matrix B alone. The meat aggregates the per-unit weighted score contributions to PSU totals and applies the stratified with-replacement formula

$$B \;=\; \sum_{h} \frac{n_h}{n_h - 1} \sum_{i \in h} (U_{hi} - \bar{U}_h)(U_{hi} - \bar{U}_h)^{\top}, \tag{3}$$

in which U_{hi} is the total of the (weighted) score contributions within PSU i of stratum h, $\bar{U}_h$ is the mean of those PSU totals over the n_h PSUs in stratum h, and the factor $n_h/(n_h - 1)$ is the with-replacement finite-PSU correction.

For the **weighted prevalence**, the linearized variable for the domain ratio estimator $\hat{p} = \sum_i w_i\delta_i y_i / \sum_i w_i\delta_i$ (with δ_i the domain indicator for a non-missing outcome and y_i the 0/1 diabetes outcome) is $z_i = w_i\,\delta_i(y_i - \hat{p})/\widehat{W}_d$, where $\widehat{W}_d = \sum_i w_i\delta_i$; its variance is obtained by substituting z_i into Equation (3). Crucially, out-of-domain units (missing outcome) are *retained* in the computation and simply contribute $z_i = 0$, so that the stratum/PSU structure of the full cohort is preserved; this is the correct domain (subpopulation) estimator and avoids the underestimation of variance that can arise from pre-deleting records and thereby corrupting the PSU structure [9, 11].

For the **survey-weighted logistic regression**, point estimates were obtained as the weighted pseudo-maximum-likelihood solution (a weighted binomial GLM), and the covariance of the coefficients was estimated with the sandwich of Equation (2). The "bread" is the weighted information matrix $A = \sum_i w_i\,p_i(1-p_i)\,x_i x_i^\top$, and the score rows entering the meat are $u_i = w_i\,(y_i - p_i)\,x_i$, with p_i the fitted probability and x_i the covariate vector. Adjusted odds ratios were obtained by exponentiating the coefficients, and 95% confidence intervals were formed on the log-odds scale using the design t-critical value with 15 degrees of freedom. Categorical predictors were dummy-coded with explicitly designated reference categories (age 20–39; BMI normal; physical activity "not meeting"; income high), and observed factor levels were validated at runtime to fail fast on any unexpected category.

For comparison, an **unweighted (naive) prevalence** was also computed, treating the sample as a simple random sample: the point estimate is the ordinary sample proportion and the 95% confidence interval uses the binomial (Wald) standard error $\sqrt{\hat{p}(1-\hat{p})/n}$ with a normal critical value. This deliberately incorrect estimator quantifies the effect of ignoring the design.

2.6 Software and reproducibility

The pipeline was implemented in Python 3.12 using `pandas`, `numpy`, and `scipy`; the weighted GLM point estimates use `statsmodels`, and figures use `matplotlib`. Two scripts reproduce the entire analysis end to end: `data_preprocessing.py` (download, merge, recode) and `survey_analysis.py` (design-based prevalence, regression, figures, and the estimates table). The exact derivation and estimation code are reproduced in Appendix A. A fixed random seed was set for determinism, although the design-based estimators are themselves deterministic.

3 Results

3.1 National prevalence: weighted versus unweighted

Among U.S. adults aged ≥ 20 years, the survey-weighted national prevalence of diagnosed diabetes was **11.85%** (95% CI: 10.70–13.01; design-based standard error 0.54%; $n_{\text{domain}} = 5{,}393$; design df = 15). The naive, unweighted estimate that treats the sample as a simple random sample was substantially higher at **16.24%** (95% CI: 15.26–17.23). Ignoring the complex survey design therefore *overstated* the national prevalence by approximately **4.4 percentage points**, a relative overestimate of about 37%. Figure 3 displays the two estimates with their 95% confidence intervals. The weighted estimate aligns closely with independently published national figures for diagnosed diabetes in this period [5, 4], whereas the unweighted estimate is biased upward because the NHANES sample intentionally overrepresents older adults and other higher-risk groups that are not present in the population in the same proportions [9, 10].

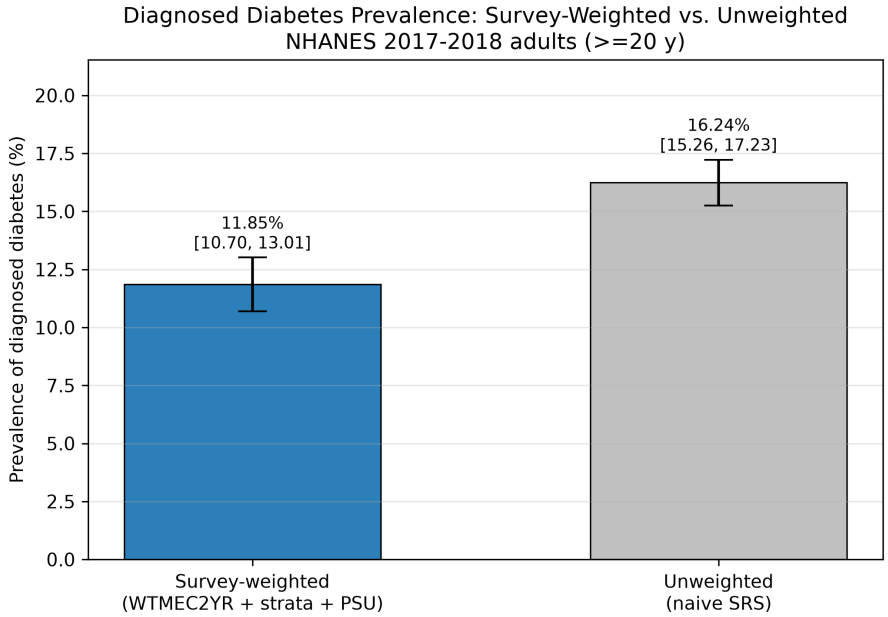

Figure 3: Diagnosed diabetes prevalence: survey-weighted vs. unweighted. Bars show the point estimate; error bars show 95% confidence intervals (design-based t_{15} interval for the weighted estimate; Wald normal interval for the naive estimate). The 4.4-percentage-point gap is entirely attributable to the survey design: weighting maps the intentionally non-representative sample composition back to the U.S. population.

It is worth noting that the two intervals do not merely shift; they are also constructed differently. The naive interval is narrower not because it is more precise in any meaningful sense but because it ignores the clustering that reduces the effective information in the sample. Even so, the dominant story here is bias in the point estimate: the weighted and unweighted intervals do not overlap, so the discrepancy cannot be explained by sampling variability alone.

3.2 Prevalence by subgroup

Figure 4 presents survey-weighted prevalence within levels of each covariate, with 95% confidence intervals. The estimates reveal steep, monotonic gradients that anticipate the multivariable model. Weighted prevalence rose from 2.2% among adults aged 20–39 to 11.6% at 40–59 and 24.5% at 60+ years; across BMI categories it climbed from 4.5% at normal weight to 10.3% among overweight and 17.3% among adults with obesity. Adults meeting the physical-activity guideline had roughly half the weighted prevalence of those not meeting it (9.0% vs. 16.9%). The income gradient was more modest and non-monotonic in the unadjusted subgroup view (low 11.6%, middle 13.6%, high 10.6%), foreshadowing the attenuated but still detectable income association that emerges after adjustment.

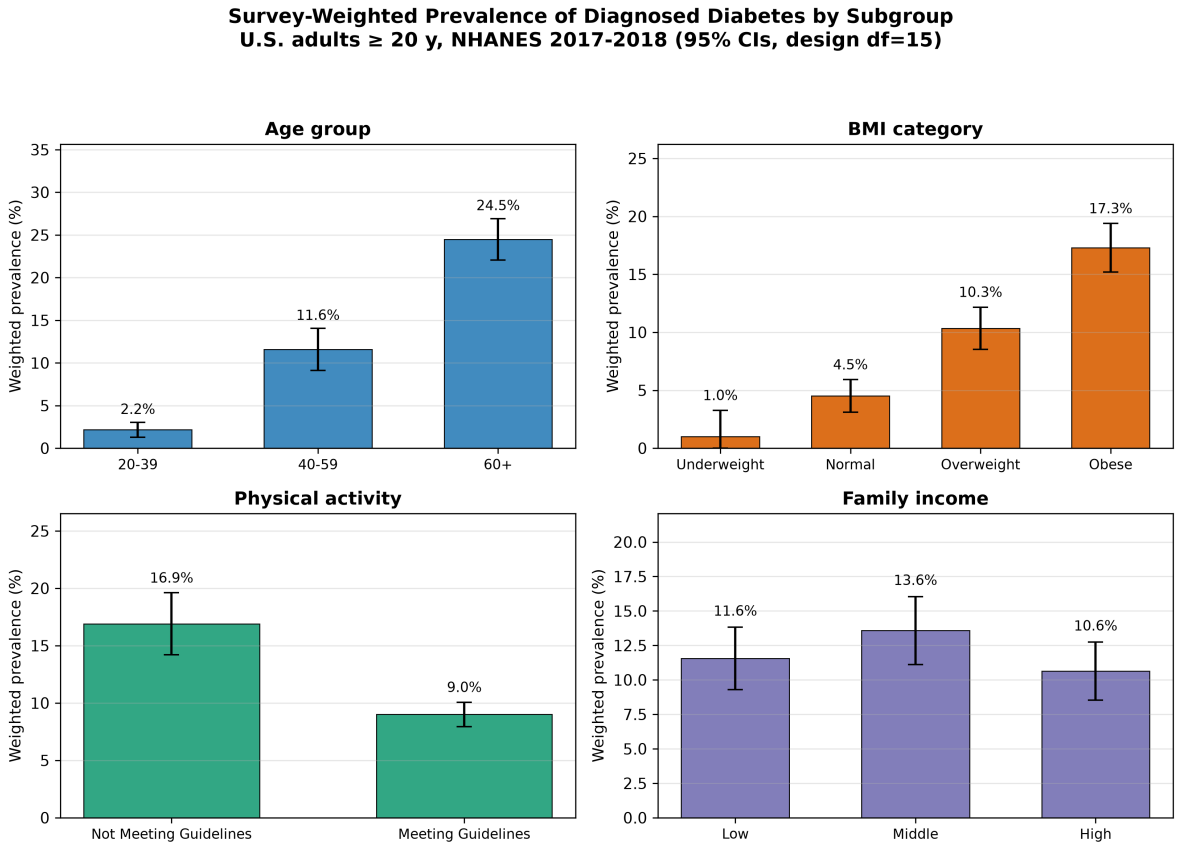

Figure 4: Survey-weighted prevalence of diagnosed diabetes by subgroup. Each panel shows weighted prevalence within levels of a covariate with 95% confidence intervals (design df = 15). The steep age and BMI gradients and the lower prevalence among those meeting the physical-activity guideline are consistent with the adjusted odds ratios in Figure 5.

3.3 Adjusted associations: survey-weighted logistic regression

The survey-weighted multivariable logistic regression (n = 4,328 complete cases) yielded the adjusted odds ratios summarized in Table 2 and displayed as a forest plot in Figure 5. Age exhibited the strongest gradient: relative to adults aged 20–39, those aged 40–59 had 5.53 times the adjusted odds of diagnosed diabetes (95% CI: 3.93–7.78) and those aged 60+ had 13.57 times the odds (95% CI: 8.44–21.80). Adiposity showed a clear dose–response: relative to normal weight, overweight was associated with an aOR of 1.99 (95% CI: 1.36–2.91) and obesity with an aOR of 4.43 (95% CI: 3.12–6.28); the underweight estimate was imprecise (aOR 0.30; 95% CI: 0.04–2.08) owing to the small number of underweight adults. Meeting the ≥150 min/week MVPA guideline was independently protective, with an aOR of 0.72 (95% CI: 0.59–0.88), corresponding to roughly 28% lower adjusted odds of diagnosed diabetes. Family income showed a graded association after adjustment: relative to high income, low income was associated with significantly higher odds (aOR 1.43; 95% CI: 1.04–1.96), while the middle-income estimate was elevated but not statistically significant (aOR 1.34; 95% CI: 0.96–1.89).

Table 2: Survey-weighted multivariable logistic regression of diagnosed diabetes on age group, measured-BMI category, physical-activity guideline status, and family-income category (n = 4,328; MEC weights + strata + PSU; design df = 15). aOR = adjusted odds ratio.

Predictor (reference)	aOR	95% CI	p-value
Age 40–59 (ref: 20–39)	5.53	3.93–7.78	< 0.001
Age 60+ (ref: 20–39)	13.57	8.44–21.80	< 0.001
BMI underweight (ref: normal)	0.30	0.04–2.08	0.207
BMI overweight (ref: normal)	1.99	1.36–2.91	0.002
BMI obese (ref: normal)	4.43	3.12–6.28	< 0.001
Meets PA guideline (ref: does not)	0.72	0.59–0.88	0.004
Income middle (ref: high)	1.34	0.96–1.89	0.084
Income low (ref: high)	1.43	1.04–1.96	0.029

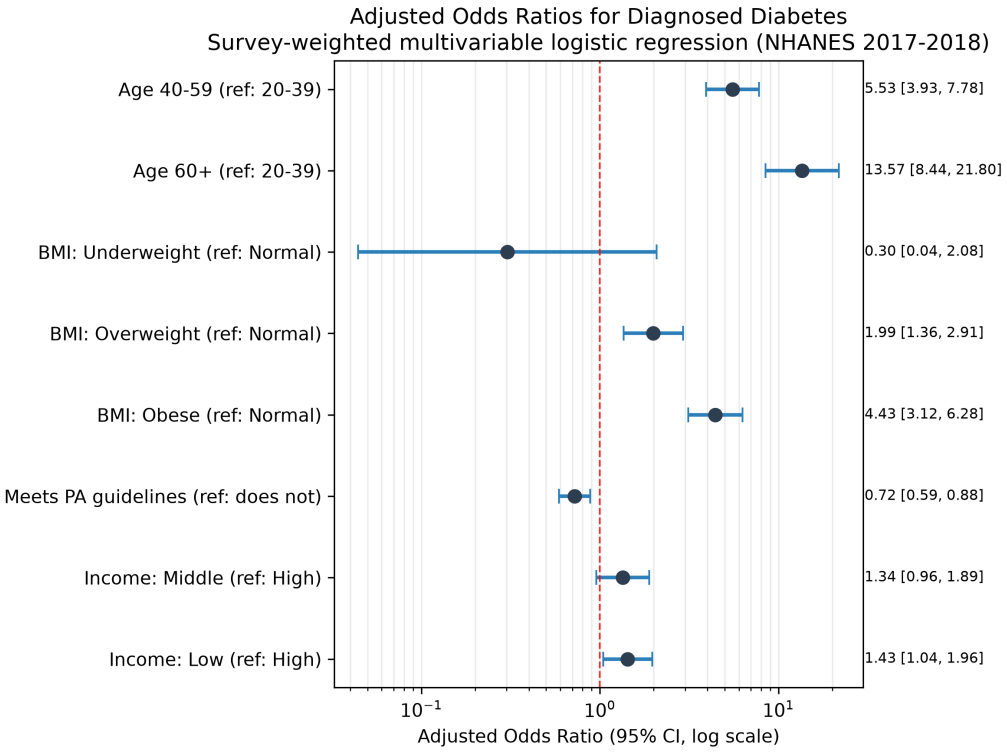

Figure 5: Adjusted odds ratios for diagnosed diabetes. Points are aORs and horizontal bars are 95% confidence intervals on a logarithmic scale; the dashed line marks the null (aOR = 1). Estimates are from the survey-weighted logistic regression with Taylor-linearization standard errors (design df = 15).

The subgroup prevalences in Figure 4 and the adjusted odds ratios in Table 2 tell a mutually consistent story. The very large unadjusted age gradient (2.2% to 24.5%) persists as the dominant adjusted effect, and the monotonic BMI gradient in the unadjusted view maps directly onto the adjusted dose–response. The full machine-readable estimates, including regression coefficients, design-based standard errors, confidence limits, and p-values, are provided in the accompanying `estimates_table.csv`.

4 Discussion

This analysis had two aims: to produce a correct, design-based estimate of diagnosed diabetes prevalence and its adjusted correlates in U.S. adults, and to demonstrate concretely what is lost when the survey design is ignored. On the first aim, the survey-weighted prevalence of 11.85% (95% CI: 10.70–13.01) is squarely consistent with the published national epidemiology of diagnosed diabetes for this period [5, 4], and the adjusted model reproduces the canonical risk architecture of type 2 diabetes. On the second aim, the side-by-side comparison shows that ignoring the weights inflated the estimate by roughly 4.4 percentage points—a difference large enough to change how the national burden would be described and, potentially, how resources would be allocated.

4.1 Why ignoring the survey design biases prevalence

The gap between the weighted and unweighted estimates is not a statistical artifact; it is a direct consequence of how NHANES is built. To guarantee adequate precision for subgroup estimates, NHANES deliberately oversamples older adults and several other higher-risk groups [9, 10]. Because diabetes prevalence rises steeply with age—from about 2.2% in the youngest adults to 24.5% in those aged 60+ in these data—an unweighted average over a sample that contains disproportionately many older adults necessarily overstates the national figure. The sampling weight `WTMEC2YR` corrects this by restoring each subgroup to its true population share, so the weighted estimate reflects the actual U.S. age (and risk) distribution rather than the engineered NHANES distribution. This is precisely the situation the survey methodology literature warns about: weighted and unweighted analyses answer different questions, and when the sampling probabilities are correlated with the outcome, the unweighted estimate is a biased estimator of the population quantity [13, 11]. A parallel, if secondary, point concerns uncertainty: treating clustered data as a simple random sample generally produces standard errors that are too small, because observations within a sampled PSU are positively correlated and therefore carry less independent information than their raw count implies [12, 11]. Design-based Taylor linearization addresses both issues at once.

4.2 Interpreting the adjusted associations

The adjusted model recovers the well-established determinants of type 2 diabetes. The steep age gradient (aOR 13.57 for 60+ vs. 20–39) reflects the cumulative, age-related decline in β-cell function and insulin sensitivity and the longer exposure to risk factors that accompanies aging. The strong, graded association with measured BMI (aOR 4.43 for obesity; 1.99 for overweight) is consistent with decades of prospective evidence that excess adiposity is the single most important modifiable driver of type 2 diabetes [19, 20, 21]; using *measured* rather than self-reported BMI strengthens this estimate by avoiding the well-known biases of self-reported height and weight. The protective association for meeting the physical-activity guideline (aOR 0.72) accords with large dose–response meta-analyses and guideline syntheses showing that regular moderate-to-vigorous activity substantially lowers the risk of type 2 diabetes and improves glycemic regulation [22, 14, 16, 23, 15]. Because this is a cross-sectional analysis, the physical-activity association should be read as consistent with—but not proof of—a causal protective effect; reverse causation, whereby diagnosed individuals reduce activity, cannot be excluded. Finally, the graded income association (aOR 1.43 for low vs. high income) mirrors the substantial socioeconomic patterning of diabetes documented in systematic reviews and in the American Diabetes Association's scientific review of social determinants of health [24, 25], reflecting pathways that include food environment, access to care, chronic stress, and health literacy.

4.3 Strengths and limitations

The principal strengths of this analysis are its correct, fully documented use of the complex survey design; its use of *measured* BMI from the MEC rather than self-report; a transparent, from-scratch implementation of the Taylor-linearization sandwich estimator that reproduces the behavior of established survey software [12]; and complete reproducibility from raw files to final estimates. Several limitations temper interpretation. First, the design is cross-sectional, so all associations are non-causal and susceptible to reverse causation, particularly for physical activity. Second, the outcome is *diagnosed* diabetes ascertained by self-report; it therefore excludes undiagnosed diabetes (which laboratory-based "total diabetes" definitions capture) and is subject to recall and awareness bias, though self-reported physician-diagnosed diabetes has generally shown high specificity in validation studies. Third, the physical-activity measure is self-reported and captures only work and recreational MVPA from the questionnaire domains; transport-related activity and measurement error may bias exposure classification toward the null. Fourth, the multivariable model was fit on complete cases ($n = 4{,}328$), and although missingness in BMI and income is modest, complete-case analysis assumes data are missing at random conditional on the model; survey-weighted multiple imputation could be explored in sensitivity analyses. Finally, a small number of extreme self-reported MVPA values were retained without truncation; capping implausible values is a reasonable sensitivity analysis but is unlikely to alter the binary guideline classification materially.

4.4 Implications

For surveillance and policy, the core message is operational: national estimates from NHANES must be computed with the sampling weights, masked strata, and masked PSUs, and their uncertainty must be quantified with a design-based variance estimator; otherwise both the point estimate and its confidence interval can be materially wrong. For prevention, the reaffirmed, independent protective association of meeting the physical-activity guideline—evident even after adjustment for age, adiposity, and income—supports continued public-health investment in physical activity as a modifiable lever against the diabetes burden [14, 22, 16], while the persistent income gradient underscores the need to address the social determinants that shape who develops and who is diagnosed with diabetes [25, 24].

5 Conclusion

Using the NHANES 2017–2018 public-use files and a correctly specified complex survey design, the weighted national prevalence of diagnosed diabetes among U.S. adults was 11.85% (95% CI: 10.70–13.01), whereas the same data analyzed without the design yielded a biased 16.24% (95% CI: 15.26–17.23). A survey-weighted multivariable logistic regression revealed strong, graded associations of diagnosed diabetes with older age and higher measured BMI, an independent protective association for meeting the $\geq$150 min/week physical-activity guideline, and higher odds among lower-income adults. Beyond reproducing the established epidemiology of type 2 diabetes, the analysis provides a concrete, fully reproducible demonstration—down to a from-scratch Taylor-linearization sandwich variance estimator—of why the NHANES survey design must be honored to obtain valid national inference.

Data and Code Availability

All source data are public-use NHANES 2017–2018 files (`DEMO_J`, `BMX_J`, `DIQ_J`, `PAQ_J`) distributed by the U.S. National Center for Health Statistics. The complete analysis pipeline (`data_preprocessing.py`,

`survey_analysis.py`, and `subgroup_prevalence.py`), the machine-readable estimates table (`estimates_table.c`
and subgroup table (`subgroup_prevalence.csv`), and all figures accompany this report and reproduce every reported number from the raw files.

References

[1] GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: A systematic analysis for the global burden of disease study 2021. *The Lancet*, 402(10397):203–234, 2023.

[2] Hong Sun, Pouya Saeedi, Suvi Karuranga, Moritz Pinkepank, Katherine Ogurtsova, Bruce B. Duncan, Caroline Stein, Abdul Basit, Juliana C. N. Chan, Jean Claude Mbanya, Meda E. Pavkov, Ambady Ramachandaran, Sarah H. Wild, Steven James, William H. Herman, Ping Zhang, Christian Bommer, Shihchen Kuo, Edward J. Boyko, and Dianna J. Magliano. IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183:109119, 2022.

[3] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda. IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, 138:271–281, 2018.

[4] Andy Menke, Sarah Casagrande, Linda Geiss, and Catherine C. Cowie. Prevalence of and trends in diabetes among adults in the united states, 1988–2012. *JAMA*, 314(10):1021–1029, 2015.

[5] Liang Wang, Xiang Li, Zhenzhen Wang, Michael P. Bancks, Mercedes R. Carnethon, Philip Greenland, Ying-Qing Feng, Hui Wang, and Victor W. Zhong. Trends in prevalence of diabetes and control of risk factors in diabetes among US adults, 1999–2018. *JAMA*, 326(8):704–716, 2021.

[6] Nadeem Sarwar, Pei Gao, Sreenivasa Rao Kondapally Seshasai, Rupa Gobin, Stephen Kaptoge, Emanuele Di Angelantonio, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies. *The Lancet*, 375(9733):2215–2222, 2010.

[7] American Diabetes Association. Economic costs of diabetes in the U.S. in 2017. *Diabetes Care*, 41(5):917–928, 2018.

[8] Michael Fang, Dan Wang, Josef Coresh, and Elizabeth Selvin. Trends in diabetes treatment and control in U.S. adults, 1999–2018. *New England Journal of Medicine*, 384(23):2219–2228, 2021.

[9] Clifford L. Johnson, Ryne Paulose-Ram, Cynthia L. Ogden, Margaret D. Carroll, Deanna Kruszon-Moran, Sylvia M. Dohrmann, and Lester R. Curtin. National health and nutrition examination survey: Analytic guidelines, 1999–2010. *Vital and Health Statistics, Series 2*, (161):1–24, 2013.

[10] Lester R. Curtin, Leyla K. Mohadjer, Sylvia M. Dohrmann, Deanna Kruszon-Moran, Lisa B. Mirel, Margaret D. Carroll, Ronette Hirsch, Vicki L. Burt, and Clifford L. Johnson. National health and nutrition examination survey: Sample design, 2007–2010. *Vital and Health Statistics, Series 2*, (160):1–23, 2013.

[11] Edward L. Korn and Barry I. Graubard. Epidemiologic studies utilizing surveys: Accounting for the sampling design. *American Journal of Public Health*, 81(9):1166–1173, 1991.

[12] Thomas Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(8):1–19, 2004.

[13] Gary Solon, Steven J. Haider, and Jeffrey M. Wooldridge. What are we weighting for? *Journal of Human Resources*, 50(2):301–316, 2015.

[14] Katrina L. Piercy, Richard P. Troiano, Rachel M. Ballard, Susan A. Carlson, Janet E. Fulton, Deborah A. Galuska, Stephanie M. George, and Richard D. Olson. The physical activity guidelines for americans. *JAMA*, 320(19):2020–2028, 2018.

[15] Fiona C. Bull, Salih S. Al-Ansari, Stuart Biddle, Katja Borodulin, Matthew P. Buman, Greet Cardon, Catherine Carty, Jean-Philippe Chaput, Sebastien Chastin, Roger Chou, Paddy C. Dempsey, Loretta DiPietro, Ulf Ekelund, Joseph Firth, Christine M. Friedenreich, Leandro Garcia, Muthoni Gichu, Russell Jago, Peter T. Katzmarzyk, Estelle Lambert, Michael Leitzmann, Karen Milton, Francisco B. Ortega, Chathuranga Ranasinghe, Emmanuel Stamatakis, Anne Tiedemann, Richard P. Troiano, Hidde P. van der Ploeg, Vicky Wari, and Juana F. Willumsen. World health organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine*, 54(24):1451–1462, 2020.

[16] Sheri R. Colberg, Ronald J. Sigal, Jane E. Yardley, Michael C. Riddell, David W. Dunstan, Paddy C. Dempsey, Edward S. Horton, Kristin Castorino, and Deborah F. Tate. Physical activity/exercise and diabetes: A position statement of the american diabetes association. *Diabetes Care*, 39(11):2065–2079, 2016.

[17] American Diabetes Association. 2. classification and diagnosis of diabetes: Standards of medical care in diabetes—2021. *Diabetes Care*, 44(Supplement 1):S15–S33, 2021.

[18] World Health Organization. Physical status: The use and interpretation of anthropometry. report of a WHO expert committee. WHO Technical Report Series 854, World Health Organization, Geneva, Switzerland, 1995.

[19] Graham A. Colditz, Walter C. Willett, Meir J. Stampfer, JoAnn E. Manson, Charles H. Hennekens, Ronald A. Arky, and Frank E. Speizer. Weight as a risk factor for clinical diabetes in women. *American Journal of Epidemiology*, 132(3):501–513, 1990.

[20] Frank B. Hu, JoAnn E. Manson, Meir J. Stampfer, Graham Colditz, Simin Liu, Caren G. Solomon, and Walter C. Willett. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine*, 345(11):790–797, 2001.

[21] Michael L. Ganz, Neil Wintfeld, Qian Li, Veronica Alas, Jakob Langer, and Mette Hammer. The association of body mass index with the risk of type 2 diabetes: A case–control study nested in an electronic health records system in the united states. *Diabetology & Metabolic Syndrome*, 6:50, 2014.

[22] Dagfinn Aune, Teresa Norat, Michael Leitzmann, Serena Tonstad, and Lars J. Vatten. Physical activity and the risk of type 2 diabetes: A systematic review and dose–response meta-analysis. *European Journal of Epidemiology*, 30(7):529–542, 2015.

[23] Sheri R. Colberg, Ronald J. Sigal, Bo Fernhall, Judith G. Regensteiner, Bryan J. Blissmer, Richard R. Rubin, Lisa Chasan-Taber, Ann L. Albright, and Barry Braun. Exercise and type 2 diabetes: The american college of sports medicine and the american diabetes association: Joint position statement. *Diabetes Care*, 33(12):e147–e167, 2010.

[24] Emilie Agardh, Peter Allebeck, Johan Hallqvist, Tahereh Moradi, and Anna Sidorchuk. Type 2 diabetes incidence and socio-economic position: A systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3):804–818, 2011.

[25] Felicia Hill-Briggs, Nancy E. Adler, Seth A. Berkowitz, Marshall H. Chin, Tiffany L. Gary-Webb, Ana Navas-Acien, Pamela L. Thornton, and Debra Haire-Joshu. Social determinants of health and diabetes: A scientific review. *Diabetes Care*, 43(11):2583–2602, 2020.

A Key derivation and estimation code

The following excerpts document the two computational cores of the analysis: the MVPA derivation (Equation (1)) and the Taylor-linearization meat matrix (Equation (3)) together with the weighted-prevalence estimator. Complete scripts accompany the report.

A.1 Weekly MVPA derivation (from `data_preprocessing.py`)

```
# Vigorous minutes are double-weighted (75 min vigorous == 150 min moderate).
work_moderate = mw_d * mw_m            # PAQ625 * PAD630
work_vigorous = 2.0 * (vw_d * vw_m)    # PAQ610 * PAD615
rec_moderate  = mr_d * mr_m            # PAQ670 * PAD675
rec_vigorous  = 2.0 * (vr_d * vr_m)    # PAQ655 * PAD660

df["mvpa_min_week"] = work_moderate + work_vigorous + rec_moderate + rec_vigorous
df["pa_guideline"]  = np.where(df["mvpa_min_week"].isna(), np.nan,
                      np.where(df["mvpa_min_week"] >= 150, 1.0, 0.0))
```

A.2 Taylor-linearization meat and weighted prevalence (from `survey_analysis.py`)

```
def linearization_meat(u, strata, psu):
    """Stratified with-replacement 'meat' B = sum_h n_h/(n_h-1) sum_i (U_hi-Ubar_h)(.)'."""
    u = np.atleast_2d(u)
    if u.shape[0] == 1 and u.shape[1] != len(strata):
        u = u.T
    k = u.shape[1]; B = np.zeros((k, k))
    df_idx = pd.DataFrame({"_s": strata, "_p": psu})
    for h, hidx in df_idx.groupby("_s").groups.items():
        hidx = np.asarray(hidx)
        psu_ids = df_idx.loc[hidx, "_p"].values
        psu_totals = np.array([u[hidx[psu_ids == p], :].sum(axis=0)
                               for p in pd.unique(psu_ids)])   # (n_h, k) PSU totals
        n_h = psu_totals.shape[0]
        if n_h < 2:
            continue
        dev = psu_totals - psu_totals.mean(axis=0, keepdims=True)
        B += (n_h / (n_h - 1.0)) * dev.T @ dev
    return B

def weighted_prevalence(df):
    w = df[WEIGHT].to_numpy(float); strata = df[STRATUM].to_numpy(); psu = df[PSU].to_numpy()
    delta = df[OUTCOME].notna().to_numpy().astype(float)      # domain indicator
    y = df[OUTCOME].fillna(0.0).to_numpy(float)
    Wd = np.sum(w * delta); p_hat = np.sum(w * delta * y) / Wd
```

```
    z = (w * delta * (y - p_hat) / Wd).reshape(-1, 1)          # linearized variable
    se = np.sqrt(linearization_meat(z, strata, psu)[0, 0])
    tcrit = stats.t.ppf(0.975, design_df(strata, psu))         # df = 30 - 15 = 15
    return p_hat, se, p_hat - tcrit*se, p_hat + tcrit*se
```

A.3 Survey-weighted logistic-regression sandwich (from `survey_analysis.py`)

```
glm  = sm.GLM(y, Xmat, family=sm.families.Binomial(), freq_weights=w).fit()
beta = glm.params; p = 1.0 / (1.0 + np.exp(-(Xmat @ beta)))
A    = (Xmat * (w * p * (1.0 - p))[:, None]).T @ Xmat          # bread: weighted information
u    = (w * (y - p))[:, None] * Xmat                           # score rows u_i = w_i(y_i-p_i)x_i
B    = linearization_meat(u, strata, psu)                      # meat
cov  = np.linalg.inv(A) @ B @ np.linalg.inv(A)                 # sandwich A^-1 B A^-1
se   = np.sqrt(np.diag(cov)); aor = np.exp(beta)               # design-based SEs and aORs
```