TASK 1

Merge Data → Filter Data → Calculate the variables/factors → Clean Data → Calculate Z-score → Model

Combine the three datasets: adult, examination, and laboratory.

To select the desired dataset, I applied several filters as outlined in the paper:

- Exclusion of Pregnant Subjects: I excluded subjects who were pregnant ("MAPF12R"), which accounted for only 5 individuals.
- Age Range: I included only subjects aged between 40-74 years.
- Race: I excluded 326 subjects who were not identified as one of the three main racial groups: non-Hispanic whites, non-Hispanic blacks, or Mexican Americans.
- Diabetes Diagnosis: I excluded 151 subjects who were diagnosed with diabetes before the age of 40, which slightly differs from the paper's exclusion of 55 subjects. This resulted in a total of 7,424 subjects (3,605 men and 3,819 women), a minor discrepancy compared to the paper's reported figures (3,609 men and 3,815 women).

Completion of Conditions as Stated in the Paper:

- Parental History of Diabetes: Subjects were classified as having a parental history of diabetes if either their biological father ("HAC5A4") or mother ("HAC5A3") had been diagnosed with the condition.
- Education Level: For the variable "HFA8R", which had 43 subjects with blank or unknown responses, I dichotomized the education levels at 12 years, considering values ranging from 1 to 17 years.
- Annual Household Income: I handled missing income data for 10 subjects with no income and 138 subjects with blank or unknown responses by dichotomizing at a threshold of $20,000.
- Smoking Status: The status of ever being a smoker was already classified in the dataset under "HAR1".
- Physical Activity: The original classification rules in the reference were based on weekly activity. However, in this dataset, physical activity was recorded monthly, requiring an adjustment to the classification rules accordingly.

Classification	Times of MET>=6 (vigorous activity)	Times of MET<6 (moderate activity)	Age
Vigorously active	>=12	—	60 up
Vigorously active	>=28	—	< 60
Moderately active	<=8	>=20	—
Lightly active	not “vigorously active” or “moderately active”		
Sedentary	engaging in no leisure-time physical activity		

And I found one question for calculating active status:

	Classification	Times of MET>=6 (vigorous activity)	Times of MET<6 (moderate activity)	Age
Case 1	Lightly active	15	27	40
Case 2	Lightly active	20	0	30

If an individual engages in vigorous activity more than 8 times but less than 28 times, and in moderate activity more than 20 times, they should not be classified as 'lightly active,' which is a lower activity level. Since they exceed 8 instances of vigorous activity, their classification should at least be 'moderately active

Anthropometric Measurements:

- Leg Length: Calculated as standing height ("BMPHT") minus sitting height ("BMPSITHT").
- Leg Length-to-Height Ratio: Calculated as the difference between standing height and sitting height divided by standing height ("BMPHT").

Target Variables Calculation:

- Body Fat: Body fat was calculated using the formula from Chumlea WC et al., with different formulas for males and females to determine Fat-Free Mass (FFM). Total Body Fat (TBF) was then calculated as weight minus FFM, and Percent Body Fat (%BF) as TBF divided by weight.
- HOMA-IR: Calculated using the HOMA2 Calculator v2.2 from the University of Oxford, requiring specific plasma glucose (3.0 to 25.0 mmol/l) and serum insulin (20 to 400 pmol/l) ranges.
- Glucose Tolerance: Subjects were first filtered to exclude those with diabetes ('HAD1'). Glucose tolerance was classified based on fasting plasma glucose and 2-hour plasma glucose levels into normal, impaired glucose tolerance (IGT), and diabetes categories.

Eligible Data for Body Fat, HOMA-IR, IGT (Glucose Intolerance):

- The eligibility for body fat calculations was based on the absence of "PEP12A1" (BIA resistance) values, affecting 751 subjects. Out of 7,424, 6,673 subjects (90%) had relevant data, slightly different from the 91% reported in the paper.
- For HOMA-IR, after excluding subjects with diabetes, filtering plasma glucose and serum insulin levels resulted in 6,142 subjects meeting the criteria from an initial 6,661.

Dataset Cleaning:

Inapplicable values for height ("BMPHT"), weight ("BMPWT"), and sitting height ("BMPSITHT") were marked as 88888 and 888888, respectively, and excluded. The leg length and its ratio to height were checked to ensure no null values existed.

Weighted Calculations:

Weighted means and standard deviations were calculated for crucial variables such as height, leg length, and leg length-to-height ratio. Z-scores for these variables were also computed.

Findings:

When modeling, data in “menstrual cycles” were uniform, like a constant value like 14 years old.

Comparison:

My results and the paper both show that anthropometric measurements correlate with metabolic health indicators. Besides, there are notable differences in the direction and magnitude of associations, particularly with height and leg length relating to body fat and diabetes. My findings suggest stronger inverse relationships between leg length-to-height ratio and HOMA-IR compared to the paper.

Discussion:

My analysis underscores the complexity of the relationship between body size and diabetes risk, suggesting that these associations may vary across different populations. There are some possible reasons for difference, like methodological differences, any discrepancies in how variables were measured or defined, or differences in the statistical models and adjustments for confounders, could account for the differing results. Also, the difference of samples and variable definitions might also cause difference, slight differences in how variables like physical activity, diabetes, or even the anthropometric measurements are defined can lead to different outcomes

TASK 2 The computed results are depicted in the subsequent table.

Contrasting disparities between the published paper and my investigation

- Sample Size Variation: Slight differences in the number of subjects included in each category may lead to statistical variations in the odds ratios and confidence intervals calculated.
- Data Processing Methods: Divergence in approaches to handling missing data, outliers, or the inclusion criteria for study participants might lead to different analytical outcomes.
- Model Specification: Variations in the statistical models used, including the covariates adjusted for in the logistic regression, can result in different odds ratios. Even minor differences in the model's specification can lead to substantial changes in the results, particularly in the width of the confidence intervals.

The comparative analysis between the findings presented in the published paper and those derived from my own investigation

Table 3 (Task1)

Anthropometric measurements	Percent body fat (Difference)			
	Published Data		My Work	
	Men	Women	Men	Women
Height				
Model 1	+0.27 (-0.07-0.61)	+0.65(0.31-0.99)*	-0.67(-0.93 – -0.41)*	-1.03(-1.36 – -0.71) *
Model 2	+0.21 (-0.13-0.56)	+0.40 (0.06-0.73)*	-0.61(-0.88 – -0.34) *	-0.68(-1.01 – -0.35) *
Model 3	–	–	–	–
Leg length				
Model 1	+0.29 (-0.37-0.62)	+0.88 (0.57-1.18)*	-0.75(-0.98 – -0.53) *	-0.86(-1.15 – -0.58) *
Model 2	+0.23(-0.11-0.57)	+0.64 (0.35-0.93)*	-0.7(-0.93 – -0.48) *	-0.65(-0.93 – -0.36) *
Model 3	–	–	–	–
Leg length-to-height ratio				
Model 1	+0.21 (-0.13-0.55)	+0.88 (0.55-1.21)*	-0.54(-0.72 – -0.35) *	-0.37(-0.59 – -0.15) *
Model 2	+0.16(-0.17-0.50)	+0.74 (0.42-1.05)*	-0.53(-0.71 – -0.35) *	-0.37(-0.58 – -0.15) *
Model 3	–	–	–	–

Table 3 (Task1)

Anthropometric measurements	Multiplicative factor		Prevalence Ratio			
	HOMA-IR		IGT	Diabetes	IGT	Diabetes
	Published Data	My Work	Published Data		My Work	
Height						
Model 1	1.00 (0.97-1.02)	1.00(0.98 – 1.01)	1.12 (1.03-1.23)*	1.19 (1.04-1.35)*	1.02(0.94 – 1.10)	1.00(0.91 – 1.09)
Model 2	0.98(0.96-1.01)	1.02(1.01 – 1.04)	1.10(1.00-1.21)	1.11 (0.96-1.29)	0.99(0.92 – 1.08)	0.97(0.88 – 1.06)
Model 3	0.97(0.95-0.99)*	1.03(1.01 – 1.05)*	1.10(0.99-1.22)	1.10 (0.94-1.29)	0.99(0.89 – 1.11)	0.97(0.85 – 1.11)
Leg length						
Model 1	1.03 (1.01-1.06)*	0.98(0.97 – 1.00)*	1.11 (1.00-1.22)*	1.25 (1.08-1.45)*	0.96(0.89 – 1.03)	0.92(0.84 – 1.00)
Model 2	1.02 (1.00-1.05)	1.00(0.99 – 1.01)	1.09 (0.98-1.21)	1.19 (1.01-1.40)*	0.94(0.87 – 1.02)	0.90(0.82 – 0.98)*
Model 3	1.00(0.98-1.02)	1.00(0.99 – 1.02)	1.09 (0.97-1.22)	1.17 (0.98-1.39)	0.93(0.84 – 1.02)	0.87(0.78 – 0.98)*
Leg length-to-height ratio						
Model 1	1.07(1.04-1.10)*	0.97(0.95 – 0.98)*	1.05 (0.95-1.16)	1.22 (1.07-1.40)*	0.88(0.82 – 0.95)*	0.83(0.76 – 0.90)*
Model 2	1.07(1.04-1.10)*	0.97(0.96 – 0.98)*	1.04 (0.94-1.16)	1.22 (1.05-1.40)*	0.89(0.83 – 0.96)*	0.84(0.77 – 0.91)*
Model 3	1.05(1.02-1.07)*	0.99(0.97 – 0.99)*	1.04 (0.94-1.16)	1.19(1.02-1.39)*	0.83(0.76 – 0.91)*	0.90(0.84 – 0.97)*

Table 2 (Task2)

Group(Odds ratio [95% CI])		Low birth weight	my analysis	High birth weight	my analysis
Model 1	All adolescents (n=1396) (n=1402)	2.01 (1.13,3.57)*	1.62 (1.08, 2.43)***	0.81 (0.48, 1.37)	0.92 (0.60, 1.39)***
Model 2		1.96 (1.12, 3.44)*	1.64 (1.09, 2.46)***	0.79 (0.46,1.35)	0.87 (0.57, 1.33)***
Model 3		1.93 (1.10, 3.38)*	1.62 (1.07, 2.44)***	0.72(0.43,1.20)	0.85 (0.55, 1.30)***
Model 4	Males (n=732) (n=734)	2.55 (1.05, 6.20)*	1.60 (0.91, 2.82)***	0.87 (0.48,1.59)	1.01 (0.62, 1.66)
Model 5		2.44 (1.02, 5.83)*	1.61 (0.91, 2.83)***	0.85 (0.46,1.58)	1.01 (0.62, 1.67)*
Model 6		2.40 (1.02, 5.67)*	1.56 (0.89, 2.76)***	0.84 (0.46, 1.56)	1.01 (0.61, 1.66)*
Model 4	Females (n=664) (n=668)	1.39 (0.62, 3.13)	1.62 (0.89, 2.93)	0.66 (0.22, 1.93)	0.71 (0.31, 1.65)
Model 5		1.38 (0.60, 3.16)	1.65 (0.91, 3.01)**	0.64 (0.21,1.96)	0.55 (0.23, 1.33)*
Model 6		1.37 (0.60, 3.13)	1.68 (0.92, 3.07)**	0.38 (0.14, 1.08)	0.48 (0.19, 1.22)**

Table 3 (Task2)

Group(Odds ratio [95% CI])	BMI Percentile (<85th percentile)n=861	my analysis n=864	BMI Percentile (≥85th percentile)n=532	my analysis n=535
Low birth weight	1.76 (0.76, 4.05)	1.54 (0.90, 2.63)***	2.13 (1.01, 4.49)*	1.73 (0.91, 3.27)**
High birth weight	1.11 (0.57, 2.18)	1.11 (0.64, 1.92)***	0.48 (0.22, 1.06)	0.67 (0.35, 1.30)**
	Waist-to-Height Ratio (≤0.5)n=893	n=895	Waist-to-Height Ratio (>0.5)n=494	n=498
Low birth weight	1.58 (0.67, 3.71)	1.50 (0.87, 2.59)***	2.35 (1.15, 4.82)*	1.82 (0.96, 3.44)**
High birth weight	0.93 (0.50, 1.71)	0.96 (0.56, 1.64)***	0.41 (0.15, 1.12)	0.75 (0.37, 1.51)*