Sampling Weights: Vitamin D Deficiency and Spine Bone Mineral Density in NHANES 2017–2018

Report 2: Code, Results, and Discussion

Due: 15 October 2025

Abstract

The association between vitamin-D deficiency and lumbar spine bone mineral density (BMD) among U.S. adults aged 50+ using NHANES 2017-2018 has been quantified in this report. Design-based analyses accounted for the complex multistage sampling (weights, strata, PSUs) and, due to BMD missingness, used complete-case subsetting with raking to restore sex, age-band, and ethnicity margins to the full sample. The estimated total number of Vitamin D deficient 50+ year-olds in the U.S population is $3,152,552$ (95% CI: $2,219,143$ - $4,085,962$). The estimated average lumbar spine bone mineral density for 50+ year-olds in the U.S. is 1.0053 g/cm^2 (95% CI: 0.9897 - 1.0210 g/cm^2). The primary adjusted Gamma log-link model estimated a small, non-significant (p = 0.1519) association between deficiency and BMD when accounting for sex, age, and ethnicity. Covariates showed expected patterns, including lower BMD among females and differences across ethnic groups, with evidence that the sex gap varies with age. The findings suggest that DXA screening for low BMD for fracture-risk reduction or other reasons should be prioritised towards higher-risk groups such as older females as opposed to vitamin-D deficient individuals.

1 Research Question

Is vitamin D deficiency associated with spinal bone mineral density among U.S. adults in NHANES 2017–2018?

2 Data Source and Variables

NHANES includes laboratory, examination, and demographic components that can be linked via the respondent sequence number `SEQN`. Public-use files also include sampling weights and design variables necessary for design-consistent estimation.

Vitamin D (VID_J.XPT)

Total 25-hydroxyvitamin D (25(OH)D) will be used as the measure of serum vitamin D to define deficiency. The variable LBXVIDMS (published April 2022) represents the sum of 25(OH)D_2 and 25(OH)D_3 measured in nmol/L of human serum. For context, human serum is blood with cells and clotting factors removed. Vitamin D_2 is produced by plants in sunlight; vitamin D_3 is produced endogenously in humans and found in animal sources. This measurement is available for 7,409 participants (males and females, all ages). From LBXVIDMS we create a binary predictor for vitamin D deficiency, defined in the literature as < 30 nmol/L.

Spinal Bone Mineral Density (DXXSPN_J.XPT)

The outcome is lumbar spine bone mineral density (BMD) for vertebrae L1–L4, stored in DXXOSBMD (published May 2020) under the Examination component. Units are g/cm^2. Exclusions for DXA include pregnancy, recent radiographic contrast, and body weight over 450 lb (DXA table limit). Scans judged invalid at any vertebra render the total spine BMD missing. This file contains data for 1,274 participants aged 50 years or older.

Demographic Variables and Sampling Weights (DEMO_J.XPT)

The demographic file contains individual and household information along with the survey design variables:

- **Weights:** full-sample 2-year Mobile Examination Center weight WTMEC2YR. When multiple components are analyzed, the appropriate weight is that of the most restrictive component (here, the DXA examination).

- **Design variables:** pseudo-primary sampling unit (SDMVPSU) and pseudo-stratum (SDMVSTRA) identifiers. These are constructed to protect confidentiality while preserving the variance structure needed for valid standard errors (two PSUs per stratum).

Demographic covariates used for confounding control and missing-data adjustment include:

- **Sex** (RIAGENDR): coded 1 = Male, 2 = Female.

- **Race/ethnicity** (RIDRETH3): categories accommodate oversampling of Asian Americans. “Mexican American” is coded 1, “Other Hispanic” 2, and non-Hispanic groups 3 (White), 4 (Black), 6 (Asian), 7 (Other).

- **Age** (RIDAGEYR): integer years to 79; age 80+ coded as 80 for disclosure protection. Although the DXA-eligible sample is 50+, RIDAGEYR is still treated as a potential confounder.

Reading datasets into R

The three datasets can be read into R using the read_xpt function from the haven package as follows.

```
library(haven)

demo <- read_xpt("DEMO_J.xpt")

dxa <- read_xpt("DXXSPN_J.xpt")

vitd <- read_xpt("VID_J.xpt")
```

3 Methods

Study Population

Adults aged $\geq$50 years with valid lumbar spine DXA and measured 25(OH)D in NHANES 2017–2018.

Exposure and Outcome

Exposure: vitamin D deficiency indicator derived from LBXVIDMS < 30 nmol/L.
Outcome: lumbar spine BMD (DXXOSBMD, g/cm^2).

Covariates

Sex, ethnicity, and age (years) will be examined as confounders within regression models as well as used for controlling for missing DXA data.

Estimation with Sampling Weights

- **Weighted estimators of totals**: national total count and prevalence of vitamin D deficiency among the DXA-eligible population, using design-based totals and averages. The functions svytotal, svymean, and svyby from the survey package in R will be used.

- **Weighted likelihood (regression)**: design-based linear models (svyglm from survey package in R) of BMD on vitamin D deficiency (unadjusted, then adjusted), with standard errors derived from the complex design.

All estimations and analysis will be performed with and without incorporation of sampling weights to compare outputs. Analysis will also be performed with and without adjustments for missing data.

Missing Data

Analysis will be conducted on missing data to determine if it should be controlled for. If there are significant differences between those with DXA data and without, the survey design object for complete cases will be adjusted to match the population margins.

4 Data Preparation

Combining datasets

The three datasets of interest can be combined using the `merge` function in `R` after first selecting the columns of interest.

```
## Merge vitamin D and DXA data columns of interest
vitd.dxa <- merge(vitd[,c("SEQN", "LBXVIDMS")],
                  dxa[,c("SEQN", "DXXOSBMD")],
                  by="SEQN")


## Select demographics columns of interest
demo.filt <- demo[,c("SEQN", "WTMEC2YR", "SDMVPSU", "SDMVSTRA",
                        "RIAGENDR", "RIDAGEYR", "RIDRETH3")]


## Merge demographics, vit D, and DXA data.
nhanes <- merge(vitd.dxa, demo.filt, by="SEQN")
```

Other preprocessing

The binary exposure variable is created using the `LBXVIDMS` variable and non-numeric variables are converted to factors.

```
## Create Vit D deficiency binary variable
nhanes$DEFICIENT <- ifelse(nhanes$LBXVIDMS < 30, 1, 0)


## Change variables to factors
nhanes$RIAGENDR <- factor(nhanes$RIAGENDR, levels = c(1,2),
                          labels = c("Male","Female"))
nhanes$RIDRETH3 <- factor(nhanes$RIDRETH3, levels = c(1:4, 6, 7),
                          labels = c("Mexican American", "Other Hispanic",
                                     "White", "Black", "Asian", "Other"))
nhanes$DEFICIENT <- factor(nhanes$DEFICIENT, levels = c(0, 1),
                              labels = c("N", "Y"))

## Create variable for banded age
nhanes$AGE_BAND <- cut(nhanes$RIDAGEYR,
                       breaks = c(seq(50, 75, 5), 80, Inf),
                        right = FALSE, include.lowest = TRUE)

## renaming variables for ease of use
names(nhanes) <- c("SEQN", "VITD", "BMD", "WEIGHT", "PSU",
                   "STRATA", "SEX", "AGE", "ETHNICITY", "DEFICIENT", "AGE_BAND")
```

```
> str(nhanes)
'data.frame': 2898 obs. of  11 variables:
 $ SEQN     : num  93705 93708 93709 93711 93713 ...
 $ VITD     : num  89.9 116 72.8 165 63.5 47.5 70.6 96.2 90.2 58.4 ...
 $ BMD      : num  1.165 0.744 1.051 1.01 0.918 ...
 $ WEIGHT   : num  8338 14372 12278 12391 166842 ...
 $ PSU      : num  2 2 1 2 1 1 2 1 2 1 ...
 $ STRATA   : num  145 138 136 134 140 147 139 143 136 139 ...
 $ SEX      : Factor w/ 2 levels "Male","Female": 2 2 2 1 1 2 1 1 2 2 ...
 $ AGE      : num  66 66 75 56 67 54 71 61 60 60 ...
 $ ETHNICITY: Factor w/ 6 levels "Mexican American",..: 4 5 4 5 3 4 6 5 1 3 ...
 $ DEFICIENT: Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ AGE_BAND : Factor w/ 7 levels "[50,55)","[55,60)",..: 4 4 6 2 4 1 5 3 3 3 ...
```

5 Exploratory Data Analysis

Dataset breakdown

By Sex

Females have a slightly higher proportion of vitamin D deficient participants and a lower proportion of missing DXA data.

Table 1: Vitamin D deficiency and DXA data missingness by sex

Sex	Number of participants	Proportion deficient	Proportion missing data
Male	1431	0.056	0.601
Female	1467	0.058	0.521

The average BMD for males is higher than for females in this dataset. The number of participants shows the number with valid total lumbar spine BMD data.

Table 2: Mean lumbar spine BMD by sex

Sex	Number of participants	Mean BMD (g/cm^2)
Male	571	1.058
Female	703	0.951

By Ethnicity

Black participants have a higher incidence of vitamin D deficiency in this dataset than any other ethnic groups. Other and White ethnicities have the highest proportion of missing DXA data. This could mean that the estimates for number of deficient individuals is over-estimated if missing data across ethnicities is not accounted for.

Table 3: Vitamin D deficiency and DXA data missingness by ethnicity

Ethnicity	Number of participants	Proportion deficient	Proportion missing data
Mexican American	331	0.059	0.508
Other Hispanic	286	0.043	0.493
White	1086	0.020	0.632
Black	696	0.130	0.550
Asian	375	0.037	0.432
Other	124	0.083	0.677

Table 4: Mean lumbar spine BMD by race/ethnicity

Race/ethnicity	Number of participants	Mean BMD (g/cm^2)
Mexican American	163	0.974
Other Hispanic	145	0.968
White	400	1.007
Black	313	1.058
Asian	213	0.934
Other	40	1.014

By Age

As age increases, the proportion of vitamin D deficient participants decreases and the proportion of missing data increases. This likely means that, without accounting for missing data across age bands, the estimated number of deficient individuals will be over-estimated.

Table 5: Vitamin D deficiency and DXA data missingness by age group

Age group	Number of participants	Proportion deficient	Proportion missing data
50-54	410	0.073	0.439
55-59	470	0.084	0.474
60-64	626	0.067	0.502
65-69	431	0.061	0.599
70-74	344	0.037	0.640
75-79	235	0.026	0.626
80+	382	0.022	0.738

Despite the highest average BMD being at the lowest age band and the lowest average BMD being at the highest age band, there is not clear evidence of a trend in the interim age bands.

Table 6: Mean lumbar spine BMD by age group

Age group (years)	Number of participants	Mean BMD (g/cm^2)
50-54	230	1.015
55-59	247	0.988
60-64	312	1.002
65-69	173	1.003
70-74	124	0.998
75-79	88	1.010
80+	100	0.961

6 Survey Design Objects

To assess how missing data and weighting adjustments affect population estimates, three survey designs are compared: the full design, the complete-case design, and the raked (adjusted) design. The full design represents the original NHANES sampling structure using all observations and weights; the complete-case design restricts analysis to respondents with non-missing values for key variables; and the raked design recalibrates the complete-case weights to match the marginal distributions of the full population. Comparing results across these designs highlights how handling missing data influences weighted estimates and ensures that subsequent analyses remain representative of the U.S. population aged 50 and over.

Full design

The full survey design object is created using the `svydesign` function, which specifies the primary sampling units (PSUs), strata, and sampling weights that define the complex survey structure.

```
des_full <- svydesign(
id = ~PSU, strata = ~STRATA, weights = ~WEIGHT,
nest = TRUE, data = nhanes
)
```

This approach ensures that all estimates and standard errors correctly reflect the multistage sampling design of NHANES, rather than treating the data as a simple random sample. Using the full dataset here captures the entire survey population before any subsetting or raking is applied.

Complete observation design

In cases where key variables contain missing data (such as BMD and VITD), we restrict analysis to respondents with complete information by subsetting the survey design object rather than the raw data. This uses a special `subset` method for `svydesign` objects, which preserves the original weights, strata, and PSU structure which ensures valid variance estimates and correct domain inference.

```
complete_design <- subset(des_full, !is.na(BMD) & !is.na(VITD))
```

Subsetting the design in this way retains all design features while excluding only cases with missing values. Operationally, this complete design represents the target population restricted to respondents aged 50+ with non-missing BMD and VITD.

Design adjusted for missing data

In order to adjust the design for missing data, the weights of the complete-cases survey design need to be adjusted to match the full-design margins. This can be done using the `rake` function. Categorical variables added previously are used for sex, age band, and race, then the population margins are calculated using the `svytable` function.

```
## population margins from the full design
pop_sex <- as.data.frame(svytable(~SEX, des_full))[, c("SEX","Freq")]
pop_age <- as.data.frame(svytable(~AGE_BAND, des_full))[, c("AGE_BAND","Freq")]
pop_race <- as.data.frame(svytable(~ETHNICITY, des_full))[, c("ETHNICITY","Freq")]

## Rake the complete cases design by the population margins
des_raked <- rake(
  design = complete_design,
  sample.margins = list(~SEX, ~AGE_BAND, ~ETHNICITY),
  population.margins = list(pop_sex, pop_age, pop_race)
)
```

The `svytable` function returns the estimated total number of individuals per group in the target population, based on the specified survey design. For example, using `svytable(~SEX, des_full)` yields totals of 52,856,895 males and 60,976,877 females (46.4% males), which matches the output obtained from the raked design. However, when using the complete-cases design, the totals decrease to 20,538,919 males and 26,438,825 females (43.7% males). This occurs because the function reports the total number of individuals (aged 50+) scaled by the proportion of valid cases. This effectively reduces counts in groups with more missing data. The lower number of males in the complete-cases design reflects the higher rate of missing data among men. Since the BMD data contain missing values, the full design cannot be used directly when referencing BMD. The raked design therefore provides a way to produce estimates consistent with the full sample proportions while still relying only on the available (non-missing) data.

7 Results

Weighted Totals

All three design objects created in the previous section will be used to estimate the total number of Vitamin D deficient 50+ year-olds in the U.S population and compared.

Crude estimate

The total number of 50+ year-old U.S. people with vitamin-D deficiency can be roughly estimated using the estimated total number of U.S. 50+ year-olds and the proportion of vitamin-D deficient individuals in the dataset. There are 5.71% deficient in the dataset and an estimated 114M 50+ year-olds in the U.S. in 2017 which equates to an estimate of approximately 6.5M deficient people. This estimate does not account for complex survey design and is therefore biased.

Full Design

Using the full survey design object, this estimate is of the total Vitamin D deficient 50+ year-olds in the U.S based on the full dataset (removing the few NA values for the Vitamin D data which are negligible compared to the NAs in the BMD data).

```
svytotal(~DEFICIENT, des_full, na.rm=T)
```

Table 7: Estimated total Vitamin D deficient based on full design

Vitamin D Status	Estimated Total	Standard Error
Not Deficient	106,595,018	4,365,546
Deficient	3,152,552	476,238

Complete-Cases Design

Using the complete-cases survey design object, this estimate is of the total Vitamin D deficient 50+ year-olds in the U.S that would be eligible for a DXA scan and not have an invalid reading. This estimate will be biased if the missingness is not at random, which there was evidence of in the exploratory data analysis.

```
svytotal(~DEFICIENT, complete_design)
```

Table 8: Estimated total Vitamin D deficient based on complete-cases design

Vitamin D Status	Estimated Total	Standard Error
Not Deficient	45,825,157	4,271,507
Deficient	1,152,587	203,328

Adjusted for Missing Values Design

Using the raked survey design object, this estimate is of the total Vitamin D deficient 50+ year olds in the U.S based on only the Vitamin D data of participants that had valid scans and scaled to match the key demographics in the full dataset. It is evident that this estimate is closer to the estimate of the full dataset compared to the complete-cases estimate. It will not be the same as the full-design result because the missing participants are still missing, the remaining ones have just been reweighted.

```
svytotal(~DEFICIENT, des_raked)
```

Table 9: Estimated total Vitamin D deficient based on adjusted design

Vitamin D Status	Estimated Total	Standard Error
Not Deficient	111,309,495	541,107
Deficient	2,524,278	541,107

The best estimate of the number of Vitamin D deficient 50+ year-olds in the U.S population is using the full design as there is no reason in this case to subset the data on the presence of a variable not used in the estimation (BMD). The reason for including this is to highlight the difference between using the different survey designs. Thus, the final estimate of the total number of Vitamin D deficient 50+ year-olds in the U.S population is $3,152,552$ with a 95% confidence interval of $2,219,143$ to $4,085,962$ obtained using the `confint` function.

Weighted Means

Weighted Mean BMD

The weighted mean BMD and standard errors using the complete-cases design and adjusted design were calculated using the `svymean` function.

Table 10: Estimated mean lumbar spine BMD based on different design objects

Design Used	Estimated Mean BMD (g/cm^2)	Standard Error
Complete-cases	1.0003	0.007
Missing-adjusted	1.0053	0.008

For comparison, the mean BMD in the original dataset is 0.9987 which is within 1 standard error of both estimates. It is slightly smaller than the estimates: this is expected due to known oversampling of Hispanic participants which have been shown to have a lower mean BMD in this dataset. The average lumbar spine bone mineral density for 50+ year-olds in the U.S. is 1.0053 g/cm^2 (95% CI: 0.9897 - 1.0210 g/cm^2).

Weighted Mean BMD by Vitamin D Deficiency Status

The `svyby` function can be used to run the `svymean` function per group in the dataset to get estimated means. Mean BMD for Vitamin D deficient and non-deficient participants has been estimated using both the complete-cases and adjusted design objects and compared to the means in the original dataset.

```
svyby(~BMD, ~DEFICIENT, complete_design, svymean)
svyby(~BMD, ~DEFICIENT, des_raked, svymean)
tapply(nhanes$BMD, nhanes$DEFICIENT, mean, na.rm=T)
```

Table 11: Lumbar spine BMD by vitamin D deficiency status under different design specifications

Design	Not deficient (N)	Deficient (Y)	P-value
Complete-cases	0.9991 (0.00731)	1.0480 (0.02222)	0.0569
Adjusted for missing data	1.0043 (0.00815)	1.0530 (0.02277)	0.0545
Original dataset means	0.9970	1.0366	0.0850

Entries are mean BMD (g/cm^2); standard errors in parentheses where available.
"Original dataset means" are simple (unweighted) sample means; SEs not shown.

For both estimates, the Vitamin D deficient group surprisingly had a higher mean BMD. T-tests have been used to determine if these differences are statistically significant using the `svyttest` function for the survey objects and the t.test function from base R for the original dataset comparison. P-values shown in Table 11. are small, however, not sufficiently small to reject the null hypothesis that there is no difference in mean BMD between the groups.

```
svyttest(BMD ~ DEFICIENT, design=complete_design)
svyttest(BMD ~ DEFICIENT, design=des_raked)
t.test(nhanes$BMD ~ nhanes$DEFICIENT)
```

Regression Models

BMD has been modelled with survey-weighted generalized linear models using a Gamma family with log-link due to BMD being strictly positive. In terms of interpretation, these models yield percent-change effects due to being on the log scale. Three nested models were fit on the raked design: an unadjusted model with vitamin D deficiency as the only predictor, an adjusted model adding sex, age, and ethnicity, and a model including a sex-age interaction. The three models are compared to each other using the anova function which use survey-corrected (Rao–Scott) likelihood-ratio tests, supporting inclusion of the covariates and the interaction.

Firstly, a non-survey-weighted generalized linear model has been fit to the data to allow us to compare the survey-weighted method, as shown below:

```
Call:
glm(formula = BMD ~ DEFICIENT, family = Gamma(link = "log"),
    data = nhanes)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002993   0.005108  -0.586   0.5580
DEFICIENTY   0.038933   0.022412   1.737   0.0826 .
```

This coefficient estimate translates to an estimated 3.97% higher BMD for vitamin D deficient individuals. This effect is not statistically significant with a p-value of 0.083, but, these standard errors and the estimate do not take into account the multi-stage survey design and so are likely biased.

Unadjusted model

```
svyglm(BMD ~ DEFICIENT, family = Gamma(link = "log"), design = des_raked)
```

The coefficient section of the output, shown below, estimates that vitamin D deficient people have `exp(0.047377) - 1 = 4.9%` higher BMD than non-deficient people, with a p-value of 0.0505 which is borderline for rejecting the null hypothesis.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.004253   0.008119   0.524   0.6086
DEFICIENTY  0.047377   0.022145   2.139   0.0505 .
```

Adjusted model

```
svyglm(BMD ~ DEFICIENT + SEX + AGE + ETHNICITY, family = Gamma(link = "log"),
 design = des_raked)
```

In the adjusted model (coefficient output below), it is evident that adding these covariates (sex, age, and ethnicity) increased the p-value for the vitamin D deficient variable coefficient. This means that some of the variation this variable was explaining in BMD in the previous model, can also be explained by either age, sex, ethnicity, or a combination of all three.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.0463663  0.0392482   1.181  0.27604
DEFICIENTY               0.0404922  0.0219082   1.848  0.10704
SEXFemale               -0.1047021  0.0173261  -6.043  0.00052 ***
AGE                     -0.0002772  0.0006471  -0.428  0.68127
ETHNICITYOther Hispanic -0.0142128  0.0194061  -0.732  0.48772
ETHNICITYWhite           0.0332759  0.0116844   2.848  0.02476 *
ETHNICITYBlack           0.0739849  0.0153485   4.820  0.00192 **
ETHNICITYAsian          -0.0384385  0.0166932  -2.303  0.05478 .
ETHNICITYOther           0.0558000  0.0289967   1.924  0.09571 .
```

The adjusted model has been compared to the unadjusted model using the `anova` function which performs a Rao-Scott likelihood ratio test between the variance in BMD explained by the nested models. The output gave a p-value of 0.001 which indicates that including these covariates explains significantly more variance in BMD than not including them in the model.

SEX-AGE Interaction model

```
svyglm(BMD ~ DEFICIENT + SEX * AGE + ETHNICITY, family = Gamma(link = "log"),
 design = des_raked)
```

The interaction between age and sex was added to the previous model. This interaction was found to be significantly associated with BMD ($p < 0.01$), showing that, for males, BMD increases with age and, for females, BMD decreases with age. Vitamin D deficiency is not significantly associated with lumbar spine BMD when accounting for these covariates ($p = 0.1519$).

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.1112844  0.0494292  -2.251  0.06532 .
DEFICIENTY               0.0360259  0.0219534   1.641  0.15190
SEXFemale                0.1867896  0.0664312   2.812  0.03068 *
AGE                      0.0022191  0.0008172   2.715  0.03486 *
ETHNICITYOther Hispanic -0.0174204  0.0188021  -0.927  0.38993
ETHNICITYWhite           0.0323341  0.0109728   2.947  0.02572 *
ETHNICITYBlack           0.0751929  0.0145066   5.183  0.00205 **
ETHNICITYAsian          -0.0398964  0.0149192  -2.674  0.03682 *
ETHNICITYOther           0.0627502  0.0349957   1.793  0.12313
SEXFemale:AGE           -0.0045623  0.0010832  -4.212  0.00561 **
```

This interaction model was compared to the previous model (Rao-Scott LRT) which yielded a p-value of 0.002. This indicates that this interaction should remain in the model. Other interactions between covariates were explored, however, none provided a significant decrease in variance explained in the outcome and also increased the complexity of the model so were not included. Using the complete-cases design object for this model was also investigated to ensure the raking was not affecting these results which yielded similar results with a p-value of 0.15487 on the deficiency variable.

8 Discussion

Principal findings. The estimated total number of Vitamin D deficient 50+ year-olds in the U.S population is $3,152,552$ (95% CI: $2,219,143$ - $4,085,962$). The average lumbar spine bone mineral density for 50+ year-olds in the U.S. is 1.0053 g/cm^2 (95% CI: 0.9897 - 1.0210 g/cm^2). No association was found between vitamin D deficiency and lumbar spine bone mineral density when accounting for the complex survey design of NHANES, as well as covariates: age, sex, and ethnicity ($p = 0.1519$).

Strengths.

- Nationally representative: NHANES complex sampling with PSUs, strata, and weights gives population-level estimates for U.S. adults aged 50+.
- Appropriate design-based inference: all estimates use `svydesign`/`svyglm`.
- Calibration to full-sample margins: raking aligns complete-case weights with sex, age-band, and ethnicity distributions from the full design, improving validity.
- Appropriate outcome model: Gamma with log link respects the strictly positive BMDs.

Limitations. Limited to 50+ year-olds in the U.S. Significant amount of missing BMD data. It also only looks at 2017-2018 and not a longer period or longitudinal effects. Vitamin-D levels could be low at the time but not yet had an effect.

Implications. The findings suggest that DXA screening for low BMD for fracture-risk reduction or other reasons should be prioritised towards higher-risk groups such as females (especially older ones) as opposed to vitamin D deficient individuals.

Future work. A few points that may be investigated further are: modelling vitamin D as a continuous variable with non-linear effects, adding non-linear effects for age, or pooling survey data from different years together following the guidelines specified on the NHANES website to strengthen the findings. Adding more covariates to analysis may provide further insights. BMI may be worthwhile to include. Low BMD may be less of a risk factor for low BMI individuals and more of a risk for those with a high BMI. Intuition suggests that high BMI may be related with both lower vitamin D (less outdoor activity) and higher BMD (a heavier load on the bones due to weight). Other covariates that may be related are activity levels (highly active people may have higher BMD) and calcium consumption or blood levels.

9 Appendix

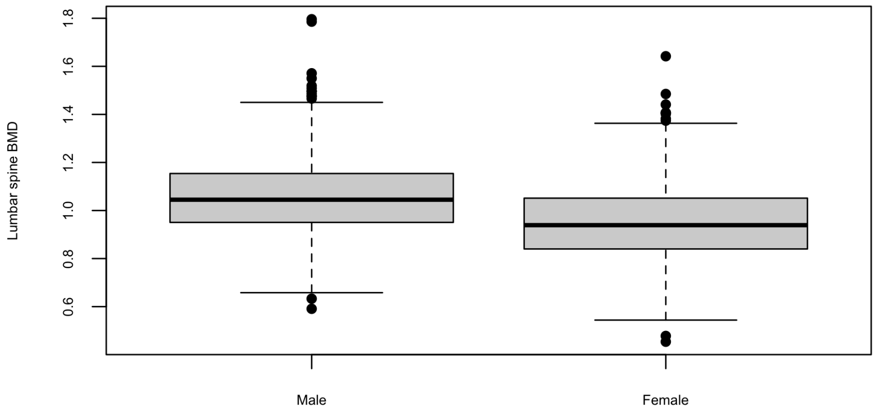

Figure 1: Boxplot of lumbar spine BMD by sex.

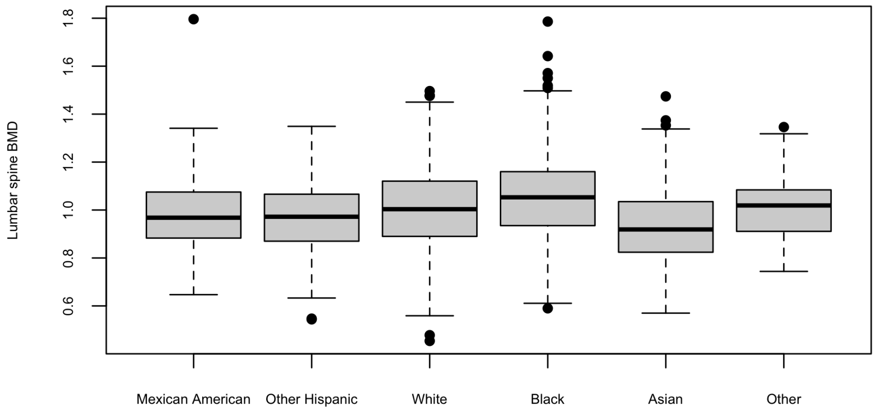

Figure 2: Boxplot of lumbar spine BMD by ethnicity.

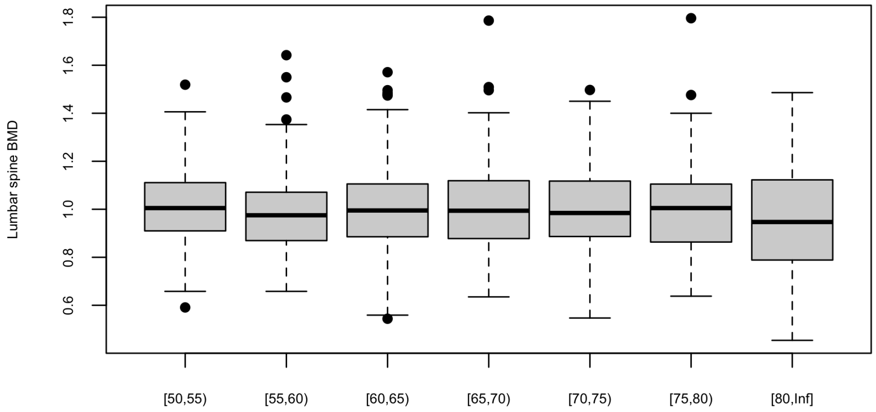

Figure 3: Boxplot of lumbar spine BMD by age band.

References

Lumley, Thomas. Complex Surveys: A Guide to Analysis Using R. Wiley, 2010.

NHANES Questionnaires, Datasets, and Related Documentation. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017. Accessed 3 Aug. 2025.

Using Survey Weights with nhanesA. https://cran.r-project.org/web/packages/nhanesA/vignettes/UsingSurveyWeights.html. Accessed 3 Aug. 2025.