STAT512: Final Project

Use of NHANES Data to Link Pesticide Exposure to Bone Mineral Density

Ryan Bontrager, Kyle McCrocklin, Shresht Venkatraman

Abstract

This exploration makes use of the NHANES (National Health and Nutritional Examination Survey) to study the association between exposure to a set of household pesticide chemicals and Bone Mineral Density (BMD). The research question was therefore "Is there a significant association between pesticide urine-biomarkers and averaged Spine-Femur Bone Mineral Density?"

Pesticide urine-biomarker levels were adjusted to ensure detection limits were consistent and log transformed to improve their distribution. The association with Average BMD was tested using several models and a final model was chosen by forward-backward stepwise selection while including 15 covariates like Age, Gender, and Family Income.

This exploration finds that a 1% increase in urine-biomarker levels of 245-trichlorophenol is associated with 0.0000024-unit reduction in Average BMD respectively. A unique finding of this exploration was that women exposed to 246-trichlorophenol were found to be associated with a larger reduction in BMD than men.

Introduction

Bone mineral density (BMD) serves as a critical indicator of skeletal health, with low BMD being associated with conditions such as osteoporosis and increased fracture risk. Understanding the environmental factors that may influence BMD, particularly pesticide exposure, is crucial for public health research and regulatory decision-making that can help with preventing low BMD induced diseases. The primary objective is to see the impact of pesticide exposure on bone mineral density across some demographic groups.

Another study has also looked at urinary chemicals and found that chemical **2,4-dichlorophenol** among others are significantly higher in farmers than non-farmers (Forté et al. 2023) and **2,4-dichlorophenol** data is also available and of interest in this study. Xu et al. (2018) looks at trends in BMD across different demographic populations and found that gender, age and race were significant in mean BMD and influenced us to include some demographic variable in the analysis.

NHANES DATA:
This study utilizes data on bone mineral density (BMD) and pesticide exposure collected during the 2005–2006 and 2007–2008 survey cycles of the National Health and Nutrition Examination Survey (NHANES) and limited to adults aged 20 or higher. The datasets from these cycles were selected due to their suitability for merging multiple variables of interest. The focus of this analysis is on identifying the relationship between environmental and demographic factors and

BMD. Before cleaning and merging the data, there were 7,700 records, which were reduced to a final dataset of 1,927 records.

The key predictor variables include demographic factors such as gender, age (in years), ethnicity, and the family poverty income ratio. Environmental factors, including the use of insect control products in the home, weed-killing products, and urinary biomarker results for specific pesticides, are also analyzed. These biomarkers include **2,5-dichlorophenol** (µg/L), **O-phenylphenol** (µg/L), **2,4-dichlorophenol** (µg/L), **2,4,5-trichlorophenol** (µg/L), and **2,4,6-trichlorophenol** (µg/L). Additionally, **urinary creatinine levels** (mg/dL) and red blood cell count (million cells/µL) are considered for their potential impact on the outcome variable. The response variables of interest are Total Femur BMD and Total Spine BMD.

In preparing the final dataset for analysis, several decisions were made to ensure consistency and accuracy. Since it was not feasible to include two response variables, the average of femur and spine BMD, both measured on the same scale (grams per square centimeter), was calculated and used as the primary response variable.

Additionally, many pesticide urine biomarker results were recorded at the lower limit of detection (LLOD). To address this, values below the LLOD were replaced with the LLOD value divided by the square root of 2, following the method outlined by Di et al. (2023) that found that many urinary chemicals did have significant effects on BMD. These adjustments were implemented to manage data limitations while preserving the integrity of the analysis.

Model Description

The first step towards identifying our best model was to investigate the predictors. The main thing we were looking for was a linear relationship between predictors and response. Scatter plots identified 5 pesticide variables whose distributions were very left skewed. These looked like good candidates for log transformation.

The log transformations did in fact improve the distribution of the variables (Figures 1,2). The following figures illustrate two such variables and their resulting log transformations. After these transformations, any relationship shown in the predictor-response scatter plots did seem linear.

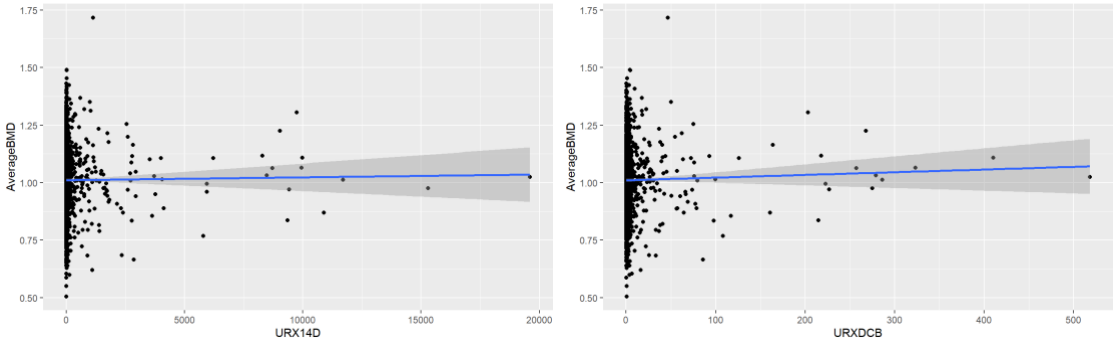

Figure 1:Original Distribution of URX14D & URXDCB variables

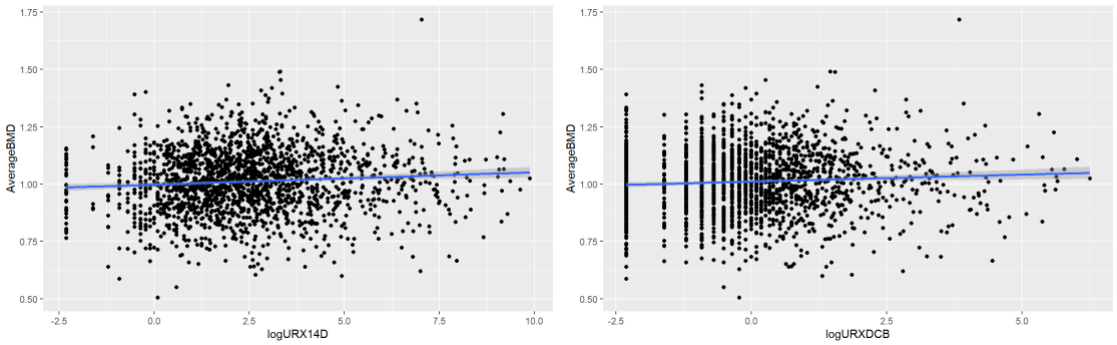

Figure 2: Log transformed URX14D & URXDCB distribution

The correlation of the variables was next considered. We found a correlation of 0.86 between the log transformed URX14D and URXDCB (Figure 3). We removed URX14D from our analysis. We originally tried including them both but ran into instability of estimated coefficients due to multicollinearity

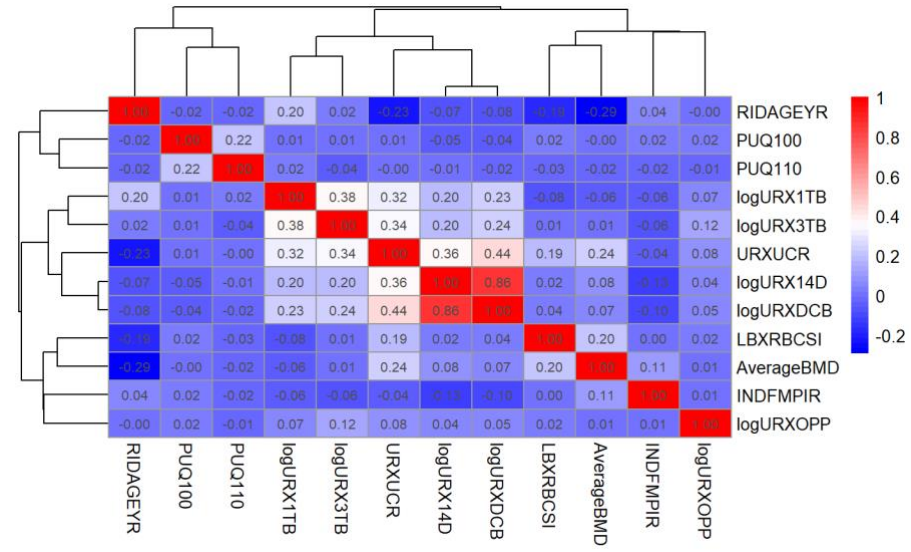

Figure 3: Correlation Matrix

A full model was fitted using all 14 first order predictors along with their interaction terms. This gives a total of 130 terms in the full model. Few of the variables were significant according to the two-tailed t-test p-values.

To determine which of the 130 terms to use in a final model, variable selection was performed. First, forward-backward stepwise selection was implemented. We used the F-test statistic, p=.05 to add, and p=.1 to drop. This resulted in 21 terms being included in the stepwise model. The effects of 15 of these were considered significant (alpha = .05).

We used the residuals vs fitted, Q-Q, and residuals vs leverage plots to look for outlier observations. It was decided to remove one observation from the dataset. The model was fit again, and the fit was improved without the outlier. This was selected as our final model shown in Table 1.

Table 1: Final Model

Covariate	NHANES Variable	# Estimate	# P-Value	Tт 95% Confidence Interval	Significance
Intercept	-	**0.987**	0.000	(0.89,1.08)	***
Participant Gender	RIAGENDR2	**0.094**	0.001	(0.04, 0.15)	**
Participant Age (yrs)	RIDAGEYR	**-0.002**	0.000	(-0.003, -0.001)	***
Participant Ethnicity 2	RIDRETH12	**-0.003**	0.788	(-0.03, 0.02)	
Participant Ethnicity 3	RIDRETH13	**-0.002**	0.833	(-0.02, 0.01)	
Participant Ethnicity 4	RIDRETH14	**0.061**	0.000	(0.04, 0.08)	***
Participant Ethnicity 5	RIDRETH15	**-0.032**	0.038	(-0.06, -0.002)	*
Poverty income ratio (PIR)	INDFMPIR	**-0.009**	0.097	(-0.02, 0.002)	.
Urinary creatinine (mg/dL)	URXUCR	**0.000**	0.814	(-0.0002, 0.0003)	
24-dichlorophenol (ug/L)	logURXDCB	**0.006**	0.347	(-0.01, 0.02)	
245-trichlorophenol (ug/L)	logURX1TB	**-0.015**	0.007	(-0.03, -0.004)	**
Red blood cell count (million cells/uL)	LBXRBCSI	**0.012**	0.069	(-0.001, 0.03)	.
Interaction Gender: Age	RIAGENDR2:RIDAGEYR	**-0.003**	0.000	(-0.003, -0.002)	***
Interaction Age: PIR	RIDAGEYR:INDFMPIR	**0.000**	0.000	(0.0002,0.001)	***
Interaction Gender: 245-trichlorophenol	RIAGENDR2:logURX1TB	**0.013**	0.085	(-0.002, 0.03)	.
Interaction Age:24-dichlorophenol	RIDAGEYR:logURXDCB	**0.000**	0.043	(-0.0005, -0.00001)	*
Interaction Age: Urinary Creatinine	RIDAGEYR:URXUCR	**0.000**	0.112	(-0.000001, 0.00001)	

The Lasso method was also used for variable selection. This came up with 16 variables to be included in the model. The R squared metric was compared between the forward-backward stepwise and the Lasso methods and the stepwise model was selected as our final model. Model diagnostics are included in the following section.

Model Diagnostics:

We believe our analysis is based on a good model. There are four key assumptions which we will be checking to confirm this: linearity, independence, homoscedasticity, and normality. After transforming our variables, the only relationships detectable between them and the predictor appeared to be linear. This satisfies our linearity assumption.

Our correlation analysis did identify strong correlation between two of the predictor variables (Figure 3) and we removed one of them to satisfy the independence assumption. Without removing one of the correlated variables, we had a multicollinearity issue where one variables coefficient was positive, and the others was negative.
Our correlation analysis did identify strong correlation between two of the predictor variables (Figure 3) and we removed one of them to satisfy the independence assumption. Without removing one of the correlated variables, we had a multicollinearity issue where one variables coefficient was positive, and the others was negative.

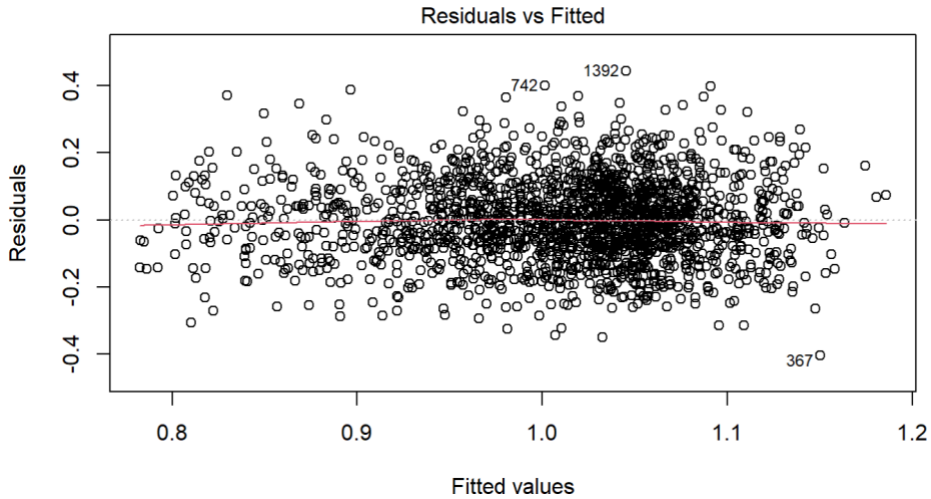

Figure 4: Residuals vs Fitted Values

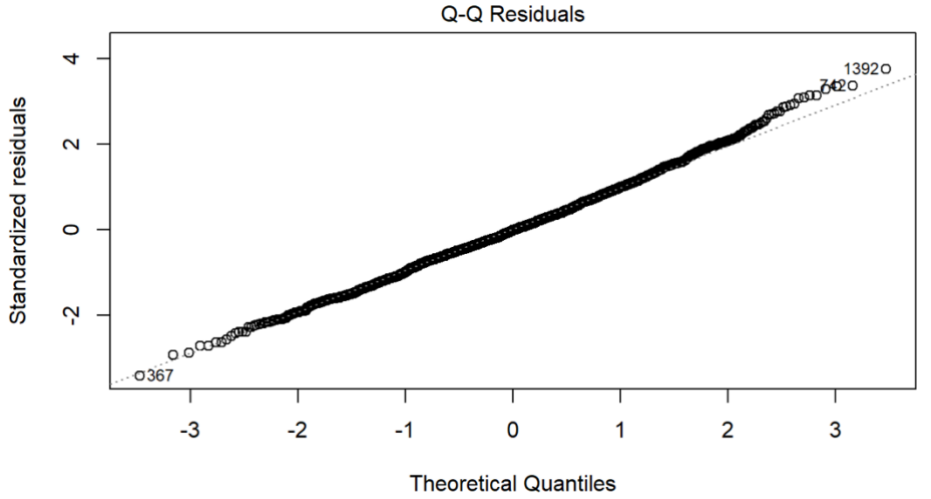

Figure 5: QQ Plot of Residuals

The residuals vs fitted plot shows that the residuals have constant variance and do not show any nonlinear pattern. This satisfies our homoscedasticity assumption. The Q-Q plot shows that the residuals are approximately normally distributed, satisfying our normality assumption. The residuals vs leverage plot does not indicate that any of the observations are outliers.

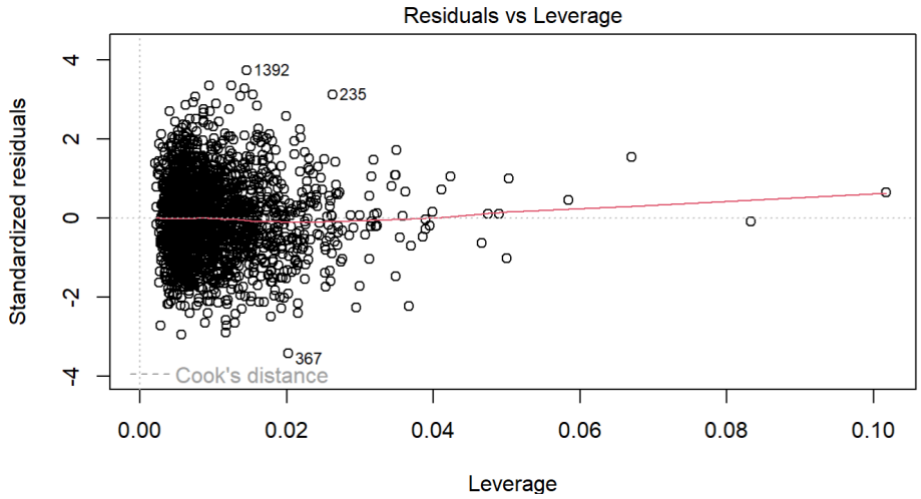

Figure 6: Residual vs Leverage Plot

Statistical Analysis

Covariate	NHANES Variable	# Estimate	# P-Value	Tт 95% Confidence Interval	Significance
Intercept	-	**0.9872**	0.000	(0.89,1.08)	***
Participant Gender	RIAGENDR2	**0.0937**	0.001	(0.04, 0.15)	**
24-dichlorophenol (ug/L)	logURXDCB	**0.0057**	0.347	(-0.01, 0.02)	
245-trichlorophenol (ug/L)	logURX1TB	**-0.0148**	0.007	(-0.03, -0.004)	**
Red blood cell count (million cells/uL)	LBXRBCSI	**0.0124**	0.069	(-0.001, 0.03)	.
Interaction Gender: Age	RIAGENDR2:RIDAGEYR	**-0.0027**	0.000	(-0.003, -0.002)	***
Interaction Age: PIR	RIDAGEYR:INDFMPIR	**0.0004**	0.000	(0.0002,0.001)	***
Interaction Gender: 245-trichlorophenol	RIAGENDR2:logURX1TB	**0.0127**	0.085	(-0.002, 0.03)	.
Interaction Age:24-dichlorophenol	RIDAGEYR:logURXDCB	**-0.0002**	0.043	(-0.0005, -0.00001)	*

Table 2: Table of relevant covariates

Table 1 showed the results of our final, fully adjusted model for the association between Average BMD and 21 covariates. We reproduce a subset of that table here with the most relevant covariates. The model found *1 pesticide with statistically significant associations* with Average BMD: **245-trichlorophenol.**

By interpreting the coefficient values in Table 1, our model shows that a 1% increase in exposure to 245-trichlorophenol in the participant's urine sample, is associated with a *decrease* in Average Bone Mineral Density of 0.00015 units. This finding in our final model is *in line with relevant literature*, as similar associations between BMD and 245-trichlorophenol have been found in other NHANES-data studies such as those by Di. et. al.

The presence of statistically significant interaction terms in the final model also yields us with some interesting implications. The coefficient value of **24-dichlorophenol** (URXDCB) did not represent a statistically significant association with Average BMD on its own, however the interaction term between this pesticide and the Age of the participant (RIDAGEYR) is statistically significant at the 0.05 level. This implies that this pesticide does not have a constant effect on average BMD but rather, it varies depending on the age of the participant. We can interpret this to mean that for every increase in age of the participant, a 1% increase in exposure to 24-

dichlorophenol is associated with a 0.0000024 unit *decrease in Average BMD*. This is an interesting in that it implies that older individuals are more vulnerable to the negative effects of this pesticide

Another interaction term that is significant at 0.1 confidence level was between GENDER2 and **245-trichlorophenol.** This term would imply that for a female participant, a 1% increase in exposure to 245-trichlorophenol is actually associated with an increase in Average BMD by 0.00013 units, offsetting the negative association that this pesticide has with males. However, since our 95% confidence interval for this interaction terms contains 0, we refrain from drawing any confident conclusions about the effects of this pesticide.

Our adjusted R2 on this final model was 0.258, suggesting that roughly 25% of the variance in the Average Bone Mineral Density of our dataset is explained by this final model.

Our Final Model Equation follows:

$$\begin{aligned}\text{AverageBMD} = {} & 0.987 - 0.002(\text{RIDAGEYR}) + 0.094(\text{RIAGENDR2}) - 0.003(\text{RIDRETH12}) \\ & - 0.002(\text{RIDRETH13}) + 0.061(\text{RIDERETH14}) \\ & - 0.032(\text{RIDRETH15}) - 0.009(\text{INDFMPIR}) + 0.00002(\text{URXUCR}) \\ & - 0.006(\text{logURXDCB}) - 0.015(\text{logURX1TB}) + 0.012(\text{LXBRCSI}) \\ & - 0.003(\text{RIAGENDR2: RIDAGEYR}) + 0.0003(\text{RIDAGEYR: INDFMPIR}) \\ & - 0.0002(\text{RIDAGEYR: logURXDCB}) + 0.000003(\text{RIDAGEYR: URXUCR}) \\ & + 0.013(\text{RIAGENDR2: logURX1TB})\end{aligned}$$

Summary of Major Findings:

This exploration sought to understand the relationship between Bone Mineral Density (BMD) and exposure to a set of 7 household pesticides (2,5-dichlorophenol, O-phenylphenol,2,4-dichlorophenol,2,4,5-trichlorophenol and 2,4,6-trichlorophenol). The dataset used in this study was the NHANES longitudinal study of participant's and contained 1,924 records of study participant's Spine and Femur Bone Density as well the exposure to the aforementioned chemicals in the participant's urine.

Our exploration only found significant relationships between Average BMD and 1 pesticide: *245-trichlorophenol*. We inferred that a 1% increase in the urine-biomarker of this pesticide was associated with a 0.0000024 unit *decrease* in BMD. Our study also found a statistically significant joint effect between Age and 24-dichlorophenol on the Average BMD of participants, implying that for every yearly increase in age, a 1% higher urine-biomarker is associated with a 0.0000024 unit *decrease in Average BMD*.

Overall, while our findings on the effects of 245-trichlorophenol are very much in line with relevant literature, this exploration also demonstrates some novel findings on the effect of this pesticide on older individuals that are innovative and unique to this study.

Suggestions & Improvements

Given more time for this study, we would have looked for better data. Without doing a deep dive of the data we picked BMD data from 2005-2008 based on other papers using data from these time periods and seeing that NHANES collected multiple questionnaires and examination datasets about pesticide use.

After trying to merge all the interested data sets together, we found that it wouldn't be possible to have all the variables originally interested in included in the analysis due to not every participant having each variable collected. We ended up choosing to reduce the number of variables to include in the overall analysis based on what would provide an adequate sample size. Given more time we would have explored different years to see if we could have had a better option of data.

Doing some research on NHANES datasets themselves we discovered that there is a restriction of the NHANES data that limits us from possibly understanding the true relationship. Which would be the inability to determine the temporal sequence of exposure and outcome, the main property of a cause-and-effect relation (LaKind et al. 2012). To effectively see the effects of pesticides on BMD we would need to see the length of time each person was exposed to each chemical. This NHANES data is just one timepoint. Some people included in the dataset may only have been exposed for a few hours and BMD may not have had enough time to be affected. We would like to run this study again but with a better collected dataset with some timepoint variables that include BMD before exposure, BMD after exposure and length of time of exposure to each chemical pesticide.

With more time, ridge regression would be another method to investigate. We did see evidence of correlation between two of the pesticide variables that applying ridge regression may help with our model and deal with any multicollinearity issues. After seeing our final model and only having two of the predictor variables having high correlation it is doubtful that ridge regression would change our best model from having a poor fit to a model with good fit.

References

1. Di, D., Zhang, R., Zhou, H. et al. *Joint Effects of phenol, chlorophenol pesticide, phthalate, and polycyclic aromatic hydrocarbon on bone mineral density: comparison of four statistical models*. Environ Sci Pollut Res 30, 80001–80013 (2023). https://doi.org/10.1007/s11356-023-28065-z

2. Forté, C., Millar, J., Colacino, J. *Integrating NHANES and Toxicity Forecaster Data to Compare Pesticide Exposure and Bioactivity by Farmwork History and US Citizenship (2023).* https://doi.org/10.1101/2023.01.24.23284967

3. LaKind, JS., Goodman, M., Naiman, DQ., *Use of NHANES Data to Link Chemical Exposures to Chronic Diseases: A Cautionary Tale*. (2012) https://doi.org/10.1371/journal.pone.0051086

4. Xu, Y., Wu, Q., *Decreasing Trend of Bone Mineral Density in US Multiethnic Population: Analysis of Continuous NHANES 2005 – 2014. (2018)*