Application of Principal Components Analysis to Urine Metal and Metalloid Exposures in the National Health and Nutrition Examination Survey (NHANES) Data

Po-Yung Cheng, Robert L. Jones, Kathleen L. Caldwell
Inorganic Radiation Analytical Toxicology Branch,
Division of Laboratory Sciences,
National Center for Environmental Health,
Centers for Disease Control and Prevention, Atlanta, GA 30333

Abstract

Human exposure to metals is an important public health issue because even low levels of exposure are associated with adverse health effects. NHANES measures metal exposures in the U.S. population; however, few researchers have examined metal co-exposures and the effects of potential interaction. To explore potential relationships between metal co-exposures, we applied principal components analysis (PCA) to 15 urine metals and metalloids in the NHANES 2013-14 national survey data. Through PCA, principal components (PC) are created from the original variables in an attempt to summarize complex data. Each PC has an eigenvalue, which indicates the variance for each PC in the data. A higher eigenvalue represents a greater variance. The first eigenvalue obtained from this analysis explained 47% of variation; the second explained 8%; the third explained 7%. The first principal component (PC1) was strongly correlated with all elements (correlation efficient (r) range: 0.56 – 0.84) except manganese. In contrast, PC2 was correlated with manganese (0.55), arsenic (-0.55) and inorganic-related arsenic species (-0.47). PC3 was correlated with barium (-0.62) and strontium (-0.48). Certain demographic characteristics, including lower income level, Asian ethnicity, female sex, and elderly, were associated with higher scores (90^{th} percentile and above) for PC1. In contrast, non-Hispanic whites with lower income level were associated with higher scores for PC2 and elderly smokers with ethnicities of non-Hispanic blacks, non-Hispanic Asians, and other Hispanics were associated with higher scores for PC3.

Key Words: principal components analysis, metal and metalloid exposure, NHANES

1. Background and motivation

Human exposure to metals such as lead (Pb), cadmium (Cd), mercury (Hg), and arsenic (As) is widespread [1]. These exposures are a critical public health concern because even relatively low levels of metals can disrupt normal development of the central nervous system, especially in early childhood [2, 3]. In general, metal exposure rarely occurs in isolation and co-exposure is likely to happen [4, 5]. Thus, metal co-exposure may pose a critical threat to human health [6]. The central nervous system is a common target organ for many environmental metals [7]. Multiple metals may interact to cause cooperative or opposing effects on neurodevelopment that are different from the main effects of exposure to each metal alone. Understanding the health effects of combinations of metals, as well as metal interactions with other chemical exposures, is important for advancing the field of environmental health and protecting human health.

In order to address potential co-exposure of metals, we applied PCA to urine metal data from NHANES 2013-14 cycle. We used this approach to summarize the interrelated metal variables by three PCs (PC1, PC2, and PC3). We further examined three PCs by assessing associations between a higher score of PCs and demographic variables.

2. Data

The Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics conducts the National Health and Nutritional Examination Survey (NHANES). The survey is a complex, multistage, probability cluster sample designed to represent the U.S. population based on age, sex, and race/ethnicity. The survey includes a physical examination and household interview, including collection of medical history, demographic, socioeconomic, and behavioral data. The survey also includes collection of biological samples for clinical chemistry tests and assessment of nutritional biomarkers and exposures to environmental chemicals.

For our study, we used urine metals data (UMS_H) which consist of a one-third subsample of persons 6 years and older from NHANES 2013-2014 cycle. We took special sample weights (WTFSM), stratification, and clustering into account while analyzing these data for logistic regression. However, due to the limitation of SAS PROC PRINCOMP procedure, stratification and clustering were not accounted for in PCA.

3. Methods

3.1 Principal Components Analysis (PCA)

PCA is an exploratory technique that helps researchers to gain a better understanding of the interrelationships between variables [8, 9, 10]. PCA is performed on a set of data with the hope of simplifying the description of a set of interrelated variables. PCA transforms the original interrelated variables into a new set of uncorrelated variables called 'Principal Components.' Each principal component is a linear combination of the original variables. The amount of information expressed by each principal component is its variance. The principal component with the highest variance is termed the 'first principal component.' The advantage of PCA is that the complexity of having a large number of interrelated variables can be reduced by utilizing only the first few principal components that explain a large proportion of the total variation.

Here are the basic concepts of PCA:

1). Assume we have a random sample of N observations for two variables, x_1 and x_2.

- Subtract the mean of each variable from each observation

  $x_1 = x_1 - \bar{x}_1$

  $x_2 = x_2 - \bar{x}_2$

- The values of x_1 and x_2 would have a mean of 0 and the sample variances, S_1^2 and S_2^2.

2). Our goal through PCA is to create two new variances C1 and C2, called principal components, that are uncorrelated.

- The new variables are linear functions of x_1 and x_2 that can be written as:
  $C1 = a_{11}x_1 + a_{12}x_2$
  $C2 = a_{21}x_1 + a_{22}x_2$
  Mean of C1 = Mean of C2 =0
  Variance of $C1 = a_{11}^2 S_1^2 + a_{12}^2 S_2^2 + 2a_{11}a_{12}Cov(S_1S_2)$
  Variance of $C2 = a_{21}^2 S_1^2 + a_{22}^2 S_2^2 + 2a_{21}a_{22}Cov(S_1S_2)$

- The variances for C1 and C2 are the first and second eigenvalues of covariance matrix of x_1 and x_2.

3). The coefficients are chosen such that:

- The variance of C1 is maximized and greater than all other variances.
- The variance of C1 ≥ the variance of C2 ≥ …
  (C1 and C2 are uncorrelated)

4). We used the SAS PROC PRINCOMP procedure with a weight statement for this study. The procedure we used did not account for stratification and clustering. To our knowledge, a modified PCA procedure capable of fully accounting for complex survey design involving stratification and clustering has not been formally developed. We are also not aware of any readily available ad hoc PCA procedure that accounts for stratification and clustering.

3.2 Logistic Regression

We used the multiple logistic regression analysis to examine characteristics of participants with PC score at the 90th percentile or higher. We adjusted the models with sex, age group, race/ethnicity, annual household income, and questionnaire questions – "Ever told you had weak/failing kidneys" and "Leak urine during nonphysical activities." Parameter estimates with p-values below an alpha (α) level of 0.05 were statistically significant. We used SUDAAN PROC RLOGIST procedure which accounted for sample weights, stratification, and clustering of complex survey data.

4. Results

4.1 Pairwise Pearson's correlations

Table 1 shows the pairwise Pearson's correlations between 15 analytes from 2517 subjects. From 105 pairs of comparison, all showed significant Pearson correlations (p-value <0.05). The moderate high correlation (0.70-0.85) occurred within 3 pairs (UBA vs. USR; UCS vs. UTL; UAS vs. Four AS).

Table 1. Correlations between 15 urine metals in NHANES 2013-24 data

	UBA	UCD	UCS	UCO	UMN	UMO	UPB	USB	USR	UTL	USN	UTU	UUR	UAS	Four_AS
UBA	1.000														
UCD	0.178	1.000													
UCS	0.451	0.479	1.000												
UCO	0.544	0.387	0.634	1.000											
UMN	0.308	0.141	0.200	0.296	1.000										
UMO	0.354	0.378	0.619	0.581	0.254	1.000									
UPB	0.404	0.561	0.595	0.495	0.223	0.466	1.000								
USB	0.339	0.373	0.496	0.473	0.296	0.482	0.530	1.000							
USR	0.756	0.282	0.561	0.634	0.258	0.492	0.485	0.394	1.000						
UTL	0.383	0.388	0.787	0.571	0.166	0.547	0.480	0.443	0.456	1.000					
USN	0.213	0.370	0.426	0.378	0.220	0.383	0.427	0.449	0.280	0.339	1.000				
UTU	0.299	0.234	0.409	0.441	0.264	0.591	0.364	0.459	0.385	0.355	0.336	1.000			
UUR	0.301	0.332	0.390	0.361	0.310	0.373	0.368	0.438	0.435	0.292	0.341	0.492	1.000		
UAS	0.240	0.324	0.520	0.368	0.113	0.423	0.392	0.323	0.360	0.471	0.266	0.294	0.280	1.000	
Four_AS	0.265	0.367	0.576	0.440	0.131	0.520	0.439	0.393	0.429	0.532	0.290	0.382	0.390	0.799	1.000

Metal Abbreviations:

UBA	Urine barium
UCD	Urine cadmium
UCS	Urine cesium
UCO	Urine cobalt
UMN	Urine manganese
UMO	Urine molybdenum
UPB	Urine lead
USB	Urine antimony
USR	Urine strontium
UTL	Urine thallium
USN	Urine tin
UTU	Urine tungsten
UUR	Urine uranium
UAS	Urine total arsenic
Four_As	Urine inorganic-related arsenic species

4.2 Eigenvalues and Scree plot

We used PCA to reduce the number of variables within a large data set. It allowed 15 urine metals to be incorporated into a fewer variables (principal components). PCA generated eigenvalues, which represent the variance for each principal component in the data. A higher eigenvalue represents a greater variance. As shown in Figure 1, PCA resulted in 3 major components. The first component (PC1) explained 47% of variation, the second component (PC2) explained 8% additional variation, and the third component (PC3) explained 7% additional variation. Therefore, PC1, PC2, and PC3 all together explain 63% of the variation among the 15 metals.

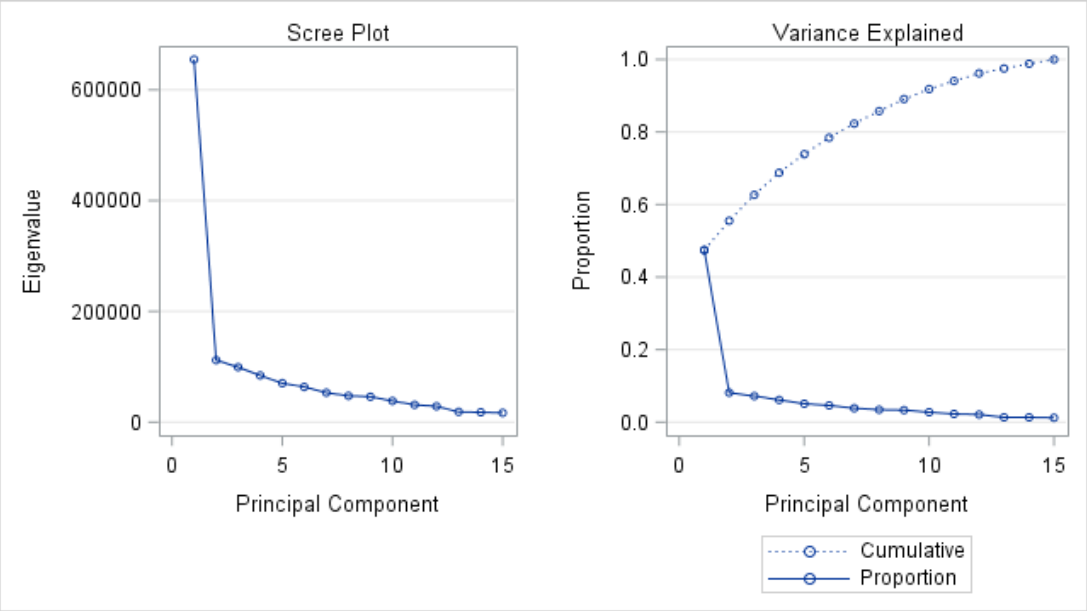

Figure 1. Eigenvalues and scree plot. The PCA generated the eigenvalues, which show the variance for each principal component in the data. A higher eigenvalue indicates a greater representation of variance. Three eigenvalues were chosen for further investigation.

4.3 Loading weights of PCs

Fig 2a, 2b, and 2c show the loading weights of each analyte PC1 and PC2, each analyte PC1 and PC3, and each analyte PC2 and PC3, respectively. The PC1 had the similar magnitude of weight (0.38-0.84) for all 15 metals with all loading weights being positive. The loading weights in PC2 were positive for UMN and negative for UAS and Four AS (Fig 2a). The loading weights in PC3 were positive for USN and negative for UBA and USR (Fig 2b). Table 2 shows the Pearson's correlations between each of the metals and PC1, PC2, or PC3.

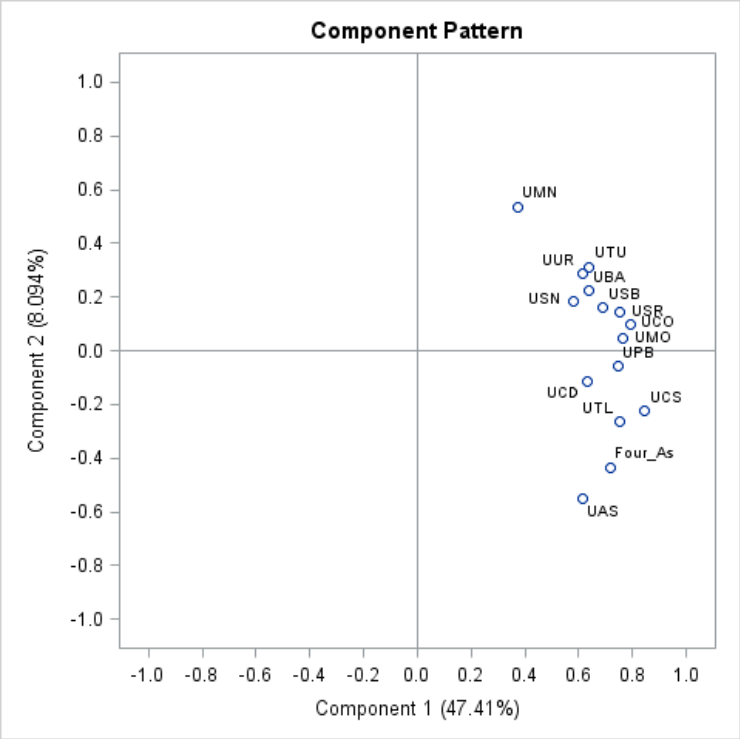

Figure 2a. Component pattern for PC1 and PC2. The pattern showed the clusters of certain metals. UMN with a higher positive loading weight was in one direction for Component 2. In contrast, UAS and Four_As were clustered in another direction with higher and negative loading weights for Component 2.

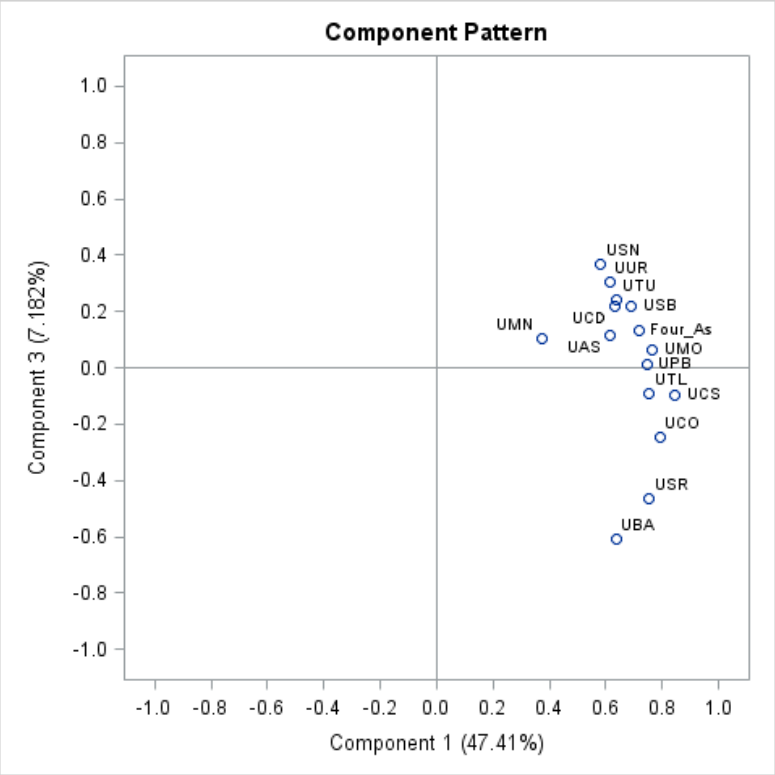

Figure 2b. Component pattern for PC1 and PC3. The pattern showed the clusters of certain metals. USN with a higher positive loading weight was in one direction for Component 3. In contrast, UBA and USR were clustered in another direction with higher and negative loading weights for Component 3.

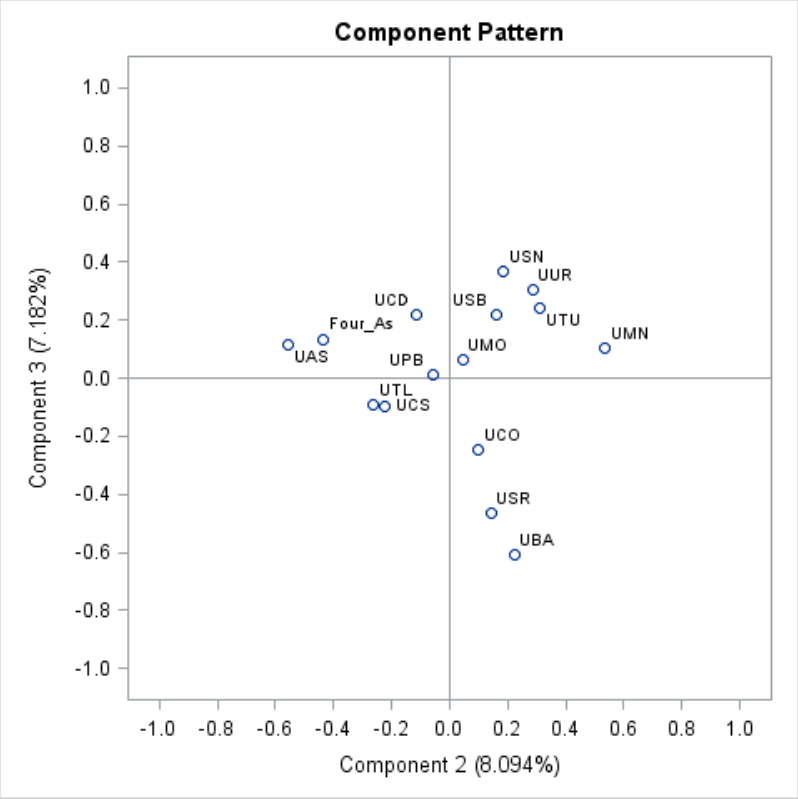

Figure 2c. Component pattern for PC2 and PC3.

Table 2. Correlations between PCs and 15 urine metals

Metal	PC 1	PC 2	PC 3
UBA	0.61	0.27	-0.62
UCD	0.58	-0.15	0.26
UCS	0.84	-0.22	-0.07
UCO	0.78	0.10	-0.25
UMN	0.38	0.55	0.06
UMO	0.76	0.02	0.08
UPB	0.73	-0.04	0.05
USB	0.69	0.18	0.23
USR	0.73	0.16	-0.48
UTL	0.74	-0.28	-0.09
USN	0.56	0.15	0.37
UTU	0.63	0.29	0.24
UUR	0.60	0.30	0.29
UAS	0.62	-0.55	0.14
Four_As	0.70	-0.47	0.16

4.4 Logistic regressions

Table 3a shows results of the multiple logistic regression which examine association between demographics and participants with a score of the PC1 at or above the 90th percentile. Age (p-value<0.001), gender (p-value<0.001), race/ethnicity (p-value<0.001), and annual household income (p-value<0.05) were statistically significant. Relative to 20-49 years old, participants aged 50 years and older were 2.75 times more likely to have a score of PC1 at or above the 90th percentile. Relative to females, males were 0.42 times less likely to have a score of PC1 at or above the 90th percentile. For race/ethnicity, compared to non-Hispanic Whites, non-Hispanic Blacks were 0.24 times less likely to have a score of PC1 at or above the 90th percentile. In contrast, non-Hispanic Asians were 5.74 times more likely to have a score of PC1 at or above the 90th percentile, relative to non-Hispanic Whites.

Table 3a. Multiple logistic regressions to examine association between demographics and participants with a score of the PC1 at or above the 90th percentile

Category	≥90 percentile
Age^	
20 - 49 year old	1.00
≥50 years old	2.75 (1.72–4.38)
Gender^	
Male	0.42 (0.28–0.61)
Female	1.00
Race/Ethnicity^	
Mexican American Americans	1.33 (0.59–3.03)
Non-Hispanic Blacks	0.24 (0.11–0.50)
Non-Hispanic Whites	1.00
Other Hispanics	1.41 (0.86–2.30)
Non-Hispanic Asians	5.74 (2.38–13.85)
Annual Household Income≠	
< $20,000	1.11 (0.45–2.76)
$20,000 to $44,999	1.95 (0.97–3.95)
$45,000 to $64,999	1.00
≥ $65,000	0.97 (0.43–2.20)
Smoker	
Yes	1.17 (0.62–2.20)
No	1.00
Ever told you had weak/failing kidneys	
Yes	1.34 (0.30–5.94)
No	1.00
Leak urine during nonphysical activities	
Yes	0.95 (0.43–2.10)
No	1.00

** odds ratio (95% confidence interval)*

*** Model was adjusted by the concentration of the urine creatinine*

ǂ p-value <0.05

p-value <0.01

^ p-value <0.001

Table 3b shows results of the multiple logistic regressions which examine association between demographics and participants with a score of the PC2 at or above the 90th percentile. Race/ethnicity (p-value<0.001) and annual household income (p-value<0.001) were statistically significant. Relative to non-Hispanic Whites, other Hispanics and non-Hispanic Asians were 0.27 and 0.32, respectively, times less likely to have a score of PC2 at or above the 90th percentile. Participants in households with an annual

income of <$20,000 and $20,000 - $44,999 were 3.23 and 1.97, respectively, times more likely to have a score of PC2 at or above the 90th percentile, relative to participants from higher-income households ($45,000-$64,999).

Table 3b. Multiple logistic regressions to examine association between demographics and participants with a score of the PC2 at or above the 90th percentile

Category	≥90 percentile
Age	
20 - 49 year old	1.00
≥50 years old	0.87 (0.65–1.16)
Gender	
Male	0.71 (0.48–1.05)
Female	1.00
Race/Ethnicity^	
Mexican American Americans	0.70 (0.36–1.38)
Non-Hispanic Blacks	0.59 (0.33–1.05)
Non-Hispanic Whites	1.00
Other Hispanics	0.27 (0.16–0.46)
Non-Hispanic Asians	0.32 (0.14–0.74)
Annual Household Income^	
< $20,000	3.23 (1.84–5.69)
$20,000 to $44,999	1.97 (1.35–2.87)
$45,000 to $64,999	1.00
≥ $65,000	1.02 (0.58–1.78)
Smoker	
Yes	0.78 (0.58–1.04)
No	1.00
Ever told you had weak/failing kidneys	
Yes	1.28 (0.35–4.60)
No	1.00
Leak urine during nonphysical activities	
Yes	1.84 (0.84–4.01)
No	1.00

** odds ratio (95% confidence interval)*

*** Model was adjusted by the concentration of the urine creatinine*

ǂ p-value <0.05

p-value <0.01

^ p-value <0.001

Table 3c shows results of the multiple logistic regressions which examine association between demographics and participants with a score of the PC3 at or above the 90th percentile. Age (p-value<0.001), race/ethnicity (p-value<0.05), smoker (p-value<0.05), and 'Ever told you had weak/failing kidneys' (p-value<0.01) were statistically significant. Relative to 20-49 years old, participants aged 50 years and older were 2.80 times more likely to have a score of PC3 at or above the 90th percentile. Relative to non-Hispanic Whites, non-Hispanic blacks, other Hispanics, and non-Hispanic Asians were 1.92, 1.86 and 2.03, respectively, times more likely to have a score of PC3 at or above the 90th percentile. Relative to non-smokers, smokers were 1.36 times more likely to have a score of PC3 at or above the 90th percentile. Relative to participants who said 'No' for the question of 'Ever told you had weak/failing kidneys', participants who said 'Yes' were 3.27 times more likely to have a score of PC3 at or above the 90th percentile.

Table 3c. Multiple logistic regressions to examine association between demographics and participants with a score of the PC3 at or above the 90th percentile

Category	≥90 percentile
Age^	
20 - 49 year old	1.00
≥50 years old	2.80 (1.80–4.36)
Gender	
Male	0.76 (0.49–1.18)
Female	1.00
Race/Ethnicity‡	
Mexican American Americans	1.04 (0.53–2.04)
Non-Hispanic Blacks	1.92 (1.01–3.67)
Non-Hispanic Whites	1.00
Other Hispanics	1.86 (1.02–3.38)
Non-Hispanic Asians	2.03 (1.03–4.01)
Annual Household Income	
< $20,000	1.61 (0.68–3.79)
$20,000 to $44,999	1.31 (0.53–3.25)
$45,000 to $64,999	1.00
≥ $65,000	0.86 (0.38–1.94)
Smoker‡	
Yes	1.36 (1.01–1.83)
No	1.00
Ever told you had weak/failing kidneys#	
Yes	3.27 (1.57–6.81)
No	1.00
Leak urine during nonphysical activities	
Yes	1.82 (0.98–3.36)
No	1.00

** odds ratio (95% confidence interval)*
*** Model was adjusted by the concentration of the urine creatinine*
ǂ p-value <0.05
p-value <0.01
^ p-value <0.001

5. Limitations

This study has certain limitations. Firstly, due to the limitation of SAS PROC PRINCOMP procedure, stratification and clustering were not accounted for in PCA. Secondly, there are various percentages of measurements below the limit of detection (LOD) in this study. LOD/(square root of 2) were imputed for those measurement results. The imputed values represent the expected values based on the assumption that the analyte had a triangular distribution of a special form in the range (0, LOD) [11]. Since the differences between the true and imputed values are likely to be small relative to the majority of values above the LOD, the influence on the data analysis should be minimal. Thirdly, the NHANES did not collect samples for younger participants (less than 6 years old). However, we expect the smaller values for this age group if we assume the urine metal exposures are similar to blood metal exposures which have the values for ages 1-5 years old. Therefore, the influence on the analysis is diminished, too.

6. Discussion

In this study, we applied PCA to NHANES 2013-14 exposure data for 15 urine metals and metalloids. This approach obtained three principal components and they together explained 63% of the variation in the data. PC1 correlated with all elements except manganese. In contrast, PC2 correlated with manganese, arsenic and inorganic-related arsenic species. PC3 correlated with barium and strontium. We also used logistic regression models to examine the association between demographics and higher scores for PC1, PC2, and PC3, respectively.

The fact that the PC1 explained only 47% of variation in the data and the correlations between PC1 and different metals were moderately high (correlation coefficient range: 0.56-0.84) indicates the PC1 was a variable for moderately expressing the amount of information for metal data. However, this fact may imply the nature of the complex co-exposure of metals in environment. Since the NHANES data we used for this study are cross sectional, the time variable is not available. Use of longitudinal data could help better elucidate the mechanisms of metal co-exposures.

The lack of function to account for stratification and clustering in SAS PROC PRINCOMP procedure may lead to spurious significance of tested associations estimated from complex survey data [11]. For example, the 95% confidence interval on the odds ratio for smokers vs. non-smokers in Table 3c was barely >1. If the same test were adjusted for stratification and clustering, the effect of smoking may become non-significant and so 95% confidence interval on the odds ratio would include 1. The advance of SAS procedure in the future may prevent this problem introduced by lacking consideration of complex survey data.

This study focused on metal exposures only. However, metal exposures could potentially interact with other chemical exposures. Currently, the NHANES measured and reported more than 300 environmental chemicals and their metabolites. By using the PCA approach, we can explore the potential interactions

between metals and these environmental chemicals. The results may improve understanding of the mechanisms of co-exposures of metals and environmental chemicals.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names and commercial sources is for identification only and does not constitute endorsement by the U.S. Department of Health and Human Services, or the U.S. Centers for Disease Control and Prevention.

References

1. CDC. Fourth National Report on Human Exposure to Environmental Chemicals, Updated Tables, March 2018
2. Malcoe LH, et al. Lead sources, behaviors, and socioeconomic factors in relation to blood lead of native american and white children: a community-based assessment of a former mining area. Environ Health Perspect. 2002; 110(Suppl 2):221–31.
3. Lidsky TI, Schneider JS. Lead neurotoxicity in children: basic mechanisms and clinical correlates. Brain. 2003; 126(Pt 1):5–19.
4. Hu H, Shine J, Wright RO. The challenge posed to children's health by mixtures of toxic waste: the Tar Creek superfund site as a case-study. Pediatr Clin North Am. 2007; 54(1):155–75.
5. Sexton K, et al. Using biologic markers in blood to assess exposure to multiple environmental chemicals for inner-city children 3–6 years of age. Environ Health Perspect. 2006; 114(3):453–9.
6. Claus Henn B, Coull BA, Wright RO. Chemical mixtures and children's health. Curr Opin Pediatr. 2014; 26(2):223–9.
7. Clarkson TW. Metal toxicity in the central nervous system. Environ Health Perspect. 1987; 75: 59-64.
8. Jolliffe, I.T. 2002. Principal Component Analysis. 2nd edition, Springer.
9. Wold, S., Esbensen, K., Geladi, P. 1987. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2: 37-52.
10. SAS/STAT 9.2 User's Guide. 2008. The PRINCOMP Procedure. SAS Institute Inc.
11. Skinner CJ., et al. The effect of sample design on principal component analysis. J Am Stat Assoc. 1986; 81(395):789-798.
12. 12. Hornung RW and Reed LD. Estimation of average concentration in the presence of non-detectable values. Appl Occup Environ Hyg. 1990; 5.1: 46-51.