Web Appendix 1: Selection of predictors to estimate blood and urine cadmium concentrations using nested predictive model

The assessment of the association of low-level exposure to cadmium, as measured in blood and urine, with peripheral arterial disease in NHANES 1999-2004 was complicated by: 1) A relatively high limit of detection (LOD) for blood cadmium (0.3 μ/L in 1999-2002 and 0.2 μ/L in 2003-2004) resulting in 20% of the study population with blood cadmium concentrations <LOD; 2) The lack of urine cadmium measurements, by design, in 2/3 of the study population with urine cadmium levels missing completely at random (MCAR). We attempted to address these two challenges with two nested prediction models. We used WinBUGS 1.4 (1, 2) to conduct joint posterior inferences on all the model parameters and missing observations given the data (See model specification details in Web Appendix 2).

First we identified potential urine and blood cadmium predictors including: 1) established strong predictors for blood and urine cadmium such as age, sex and smoking (3), and 2) factors associated with peripheral arterial disease (race, low education, body mass index, blood lead, c-reactive protein, total cholesterol, HDL-cholesterol, cholesterol lowering medication use, systolic blood pressure, blood pressure lowering medication use, diabetes, and glomerular filtration rate). Second, we used a backward stepwise process using linear regression separately for log-transformed blood and urine cadmium in smoking status strata. We selected predictors with smaller standard errors and statistically significant coefficients to maximize the accuracy of the predictive model while minimizing error. Age, sex menopause and survey year were systematically included in all the models.

For the stepwise process, blood cadmium observations <LOD were initially imputed as the LOD divided by the square root of 2 (4). All models were fit with complete data sets (i.e., each participant had information in all covariates except for urine cadmium that was missing at random in 2/3 of the sample). Web Table 1 provides a summary of the finally selected predictors that were used in the two nested predictive equations. For the final model, smoking status was included as an interacting covariable, and all the variables selected from the smoking strata were pooled together in the same model (See Web Appendix 2).

Web Table 1. β coefficients (standard errors) for predictors of log-transformed blood and urine cadmium by smoking status selected through a backward stepwise process

	Never Smoker		Former Smoker		Current Smoker	
	Blood (R^2=0.34)	**Urine (R^2=0.58)**	**Blood (R^2=0.47)**	**Urine (R^2=0.72)**	**Blood (R^2=0.42)**	**Urine (R^2=0.68)**
Urine Cadmium Log(µg/L)	0.25 (0.07)	--	0.47 (0.04)	--	0.43 (0.07)	--
Urine Creatinine Log(mg/dL)	-0.28 (0.08)	0.90 (0.07)	-0.40 (0.06)	0.95 (0.04)	-0.45 (0.07)	0.90 (0.04)
Blood Cadmium Log(µg/L)	--	0.57 (0.05)	--	0.65 (0.04)	--	0.43 (0.05)
Sex § 0-Men; 1-Women	0.15 (0.04)	0.32 (0.07)	0.10 (0.07)	0.27 (0.10)	0.25 (0.09)	0.17 (0.10)
Age § Years	0.004 (0.001)	0.009 (0.002)	0.004 (0.002)	0.01 (0.002)	-0.01 (0.003)	0.02 (0.002)
Menopause § 1-Yes; 0-No	-0.005 (0.002)	0.13 (0.08)	-0.07 (0.07)	-0.01 (0.09)	-0.004 (0.002)	0.006 (0.12)
Survey year 2001-2002 § 1-Yes; 0-No	-0.07 (0.06)	0.17 (0.05)	-0.12 (0.06)	0.06 (0.07)	-0.13 (0.08)	0.15 (0.11)
Survey year 2003-2004 § 1-Yes; 0-No	-0.16 (0.06)	0.17 (0.06)	-0.28 (0.07)	0.17 (0.05)	0.02 (0.09)	-0.02 (0.10)
Race 1-Other* / Black**; 0-Else	0.14 (0.05) *		--	--	--	-0.16 (0.07) **
Cotinine Log(ng/mL)	--	--	0.04 (0.02)	--	0.11 (0.03)	0.08 (0.02)
Lead Log(µg/dL)	0.16 (0.03)	--	--	--	0.22 (0.06)	--
Cholesterol-lowering med. 1-Yes; 0-No	-0.06 (0.03)	0.15 (0.04)	--	--	--	
Diabetes 1-Yes; 0-No	-0.10 (0.05)	--	-0.17 (0.06)	--	--	--
HDL-cholesterol mg/dL	--	--	--	--	0.005 (0.002)	--
Estimated GFR mL/min/1.72m^2	--	0.003 (0.001)	-0.003 (0.001)	0.003 (0.001)	-0.004 (0.002)	-0.02 (0.005)
Body Mass Index Kg/m^2	-0.005 (0.002)	--	--		--	
Systolic Blood Pressure mm Hg	0.002 (0.0007)	-0.002 (0.001)	0.002 (0.001)		--	--

§ Age, sex, menopause and survey year were systematically included in the models regardless of the selection process

REFERENCES

1. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing.* 2000;10:325-337.

2. Spiegelhalter D, Thomas A, Best A, Lunn D. WinBUGS Version 1.4 User Manual, MRC Biostatistics Unit, Cambridge. 2003. (www.mrc-bsu.cam.ac.uk/bugs/). (Accessed February, 2010).

3. Nordberg GF, Nogawa K, Nordberg M, Friberg L. Cadmium. In: Nordberg GF, Fowler BF, Nordberg M, Friberg L, eds. *Handbook on the toxicoloty of metals.* Amsterdam: Elsevier; 2007:445-486.

4. Hornung RW, Reed LD. Estimation of Average Concentration in the Presence of Nondetectable Values. *Applied occupational and environmental hygiene.* 1990;5(1):46-51.

Web Appendix 2: Nested predictive model formulation using Markov Chain Monte Carlo (MCMC) by Gibbs sampling

Below we provide the technical details for the implementation of the Bayesian inference for the joint prediction models for missing data. In our model $W_{b,i}$ and $W_{u,i}$ denote the log of the observed blood and urine cadmium concentration for subject i, respectively. For blood concentrations, measured values were unobserved (set to missing) if their values were below the limit of detection (LOD). The LOD was 0.3 for NHANES 1999-2002 and 0.2 for NHANES 2003-2004. For urine cadmium, a random subset of 2/3 of participants had, by design, no measured urine cadmium levels (Missing Completely at Random or MCAR). If $w_{b,i}$ and $w_{u,i}$ denote the observed blood and urine cadmium concentrations, respectively, then the structure of the data for blood and urine cadmium is as follows:

Blood Cadmium

$W_{b,i}$ =NA if NHANES 1999-2002 and the observation is below LOD = log(0.3)
$W_{b,i}$ =NA if NHANES 2003-2004 and the observation is below LOD = log(0.2)
$W_{b,i} = w_{b,i}$ if observation is above the corresponding LOD in either period

Urine cadmium

$W_{u,i}$ =NA if subject i was not selected for urine cadmium sampling
$W_{u,i} = w_{u,i}$ if subject i was selected for urine cadmium sampling

"NA" stands for "Not Available". We assume that the log-concentrations of blood and urine cadmium follow normal distributions, that is: Normal($\mu_{b|u,i}$, $\sigma_{b|u}^2$) and Normal($\mu_{u,i}$, σ_u^2).

We model the joint distribution of ($W_{b,i}$, $W_{u,i}$) as the product of the conditional distribution of blood cadmium concentration given urine cadmium concentration, denoted by [$W_{b,i}|W_{u,i,}$ parameters$_{b|u}$], and the marginal distribution of the urine cadmium concentration, denoted by [$W_{u,i,}$ parameters$_u$]. The joint distribution is then

$$[W_{b,i}, W_{u,i} \mid \text{parameters}_{b,u}] = [W_{b,i}|W_{u,i,}\ \text{parameters}_{b|u}]*[\ W_{u,i,}\ \text{parameters}_u] = \text{Normal}(\mu_{b|u,i}, \sigma_{b|u}^2) * \text{Normal}(\mu_{u,i}, \sigma_u^2)$$

The conditional mean $\mu_{b|u,i}$ uses both urine cadmium and other predictors as follows

$$\mu_{b|u,i} = \beta_u W_{u,i} + \beta_{fs,u} W_{u,i}\ I(\text{frm. smk.},i) + \beta_{cs,u} W_{u,i}\ I(\text{curr. smk.},i) + \mathbf{X}_i\boldsymbol{\beta}.$$

Here β_u is the main effect coefficient of the urine cadmium, $\beta_{fs,u}$ is the interaction coefficient between urine cadmium and being a former smoker, $\beta_{cs,u}$ is the interaction coefficient between urine cadmium and being a current smoker, and $\mathbf{X}_i$ is a row vector containing subject specific predictors of blood cadmium. We used the following additional blood cadmium predictors: age, sex, smoking, menopause, survey year, urine creatinine, race, serum cotinine, blood lead, cholesterol lowering medication intake, diabetes status, serum HDL-cholesterol, estimated GFR, body mass index and systolic blood pressure.

The marginal mean of the urine cadmium distribution, $\mu_{u,i}$, does not contain blood cadmium as a predictor. Because we use a joint model for blood and urine cadmium, the conditional mean of urine cadmium

concentration given blood concentration implicitly depends on blood concentration, even though the marginal mean does not. More precisely,

$$\mu_{u,i} = \mathbf{Z}_i\boldsymbol{\gamma},$$

where $\mathbf{Z}_i$ is a row vector containing subject specific predictors of urine cadmium. We used the same additional predictors except: blood lead, diabetes status, serum HDL-cholesterol, and body mass index. The conditional standard deviation, $\sigma_{b|u}$, and the marginal standard deviation, σ_u, are assumed to be constant in the population.

Our model uses Markov Chain Monte Carlo (MCMC) by Gibbs sampling simulations (1, 2) to obtain the joint posterior distribution of all parameters and missing observations given the data and the joint model for blood and urine cadmium. The joint posterior distribution has three major components: 1) the mean value of $\mu_{b|u,i}$ and $\mu_{u,i}$ calculated for each subject; 2) the variance of the errors $\sigma_{b|u}^2$ and σ_u^2 and 3) the missing values in $W_{b,i}$ (truncated at the LODs) and $W_{u,i}$ (missing completely at random due to sampling restrictions). For all these components thousands of simulations from their posterior distribution are obtained. Particularly, the entire distribution of each below LOD blood and MCAR urine cadmium observations is obtained.

Our approach is to simultaneously estimate the model parameters and simulate missing data by using a method that is similar in nature to a multiple imputation approach with a few important differences. First, we use a simulation approach that correctly incorporates the uncertainty in parameter estimation and missing data prediction. Second, we simulate thousands of complete data sets, instead of 2 or 10, and use the rules of probability to automatically pool results in the joint posterior analysis. Third, we incorporate tobit-like models with standard imputation modeling to account for missing due to below LOD when the LOD varies in the sample.

A MCMC simulation method by Gibbs sampling (1, 2) was preferred for this paper for the following reasons: 1) it is flexible by allowing the specification of models that are necessarily complex to represent the complex nature of the data; 2) it is computationally feasible and 3) it correctly incorporates measurement error and its effects on parameters estimation (3, 4). Our code would require minimal changes to include random effects, smoothing components and miss-measured data (see e.g., Crainiceanu, Ruppert, and Wand (3), and Crainiceanu and Goldsmith (5), for a deeper discussion on Bayesian methods and more advanced examples). Nonetheless, exploration of other imputation models is beyond the scope of this paper.

We provide the BUGS code for reproducibility purposes:

```
model {#Begin model

#Likelihood of the model
     for (i in 1:NOBS) {#Begin loop over i

     #Establishing the observed range for blood cadmium.
     #c is a vector containing the two log-transformed LOD
     #svy3 is an indicator for survey wave 2003-2004

     upper[i]<-c[svy3[i]+1]*is.below[i]+
               upperlim*(1-is.below[i])


     #Model for observed log Blood Cadmium levels
     w1[i]~dnorm(m_1[i],tau_m1)I(-9,upper[i])
     m_1[i]<-beta0+beta.u*w2[i]+
             beta.fs.u*frm.smk[i]*w2[i]+
             beta.cs.u*curr.smk[i]*w2[i]+
             beta[1]*female[i]+beta[2]*age[i]+
             beta[3]*race.4[i]+beta[4]*cholmed[i]+
             beta[5]*menop[i]+beta[6]*diab[i]+
             beta[7]*loglead[i]+beta[8]*log.u.creat[i]+
             beta[9]*logcot[i]+beta[10]*svy2+
             beta[11]*svy3[i]+beta[12]*frm.smk[i]+
             beta[13]*curr.smk[i]+beta[14]*bmi[i]+
             beta[15]*hdl.chol[i]+beta[16]*gfr[i]+
             beta[17]*systolicbp[i]

     #Model for observed log Urine Cadmium levels
     w2[i]~dnorm(m_2[i],tau_m2)
     m_2[i]<-gamma0+gamma[1]*female[i]+gamma[2]*age[i]+
             gamma[3]*race.2[i]+gamma[4]*cholmed[i]+
             gamma[5]*menop[i]+gamma[6]*log.u.creat[i]+
             gamma[7]*logcot[i]+gamma[8]*svy2+
             gamma[9]*svy3[i]+gamma[10]*frm.smk[i]+
             gamma[11]*curr.smk[i]+gamma[12]*gfr[i]
             gamma[13]*systolicbp[i]

     }#End loop over i


#Define the uninformative priors
     tau_m1~dgamma(0.001, 0.001)
     tau_m2~dgamma(0.001, 0.001)
     beta0~dnorm(0,1.0E-6)
     beta.u~dnorm(0,1.0E-6)
     beta.fs.u~dnorm(0,1.0E-6)
     beta.cs.u~dnorm(0,1.0E-6)
     gamma0~dnorm(0,1.0E-6)

     for (k in 1:K){
```

```
      beta[k]~dnorm(0,1.0E-6)
      }

      for (l in 1:L){
      gamma[l]~dnorm(0,1.0E-6)
      }

}#End model
```

Data

Data consists on the log transformed blood (`w1`) and urine (`w2`) cadmium variables and its corresponding predictors (`female, age, race.2, race.4, cholmed, menop, diab, loglead, log.u.creat, logcot, svy2, svy3, frm.smk, curr.smk, bmi, systolicbp, hdl.chol, gfr`), sample size (`NOBS`), a vector containing limits of detection for surveys 1999-2002 and 2003-2004 (`c=c(log(0.3),log(02))`), a below the detection limit indicator vector (`is.below`), a very large log-transformed upper limit of observations (`upperlim=1000`) and number of parameters for the blood and urine cadmium predictive equations (excluding the constants, urine cadmium and interaction coefficients) (`K=17, L=13`).

Initial values

Initial values were provided for the blood and urine cadmium marginal precisions τ_b and τ_u (`tau_m1, tau_m2`), predictive equations coefficients (`beta0, beta.u, beta.fs.u, beta.cs.u, beta[], gamma0, gamma[]`), and observed log-transformed urine and cadmium levels (`w1, w2`).

Both data and initial values are specified and processed in R-software (6) and then used in WinBUGS through the `bugs()` function implemented in the R2WinBUGS package (7). R-software was also used to do output checking and processing as well as plotting.

REFERENCES

1. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing.* 2000;10:325-337.

2. Spiegelhalter D, Thomas A, Best A, Lunn D. WinBUGS Version 1.4 User Manual, MRC Biostatistics Unit, Cambridge. 2003. (www.mrc-bsu.cam.ac.uk/bugs/). (Accessed February, 2010).

3. Crainiceanu CM, Ruppert D, Wand MP. Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software.* 2005;14(14):1-24.

4. Crainiceanu CM, Ruppert D, Carroll RJ, Adarsh J, Goodner B. Spatially Adaptive Penalized Splines with Heterosdecastic Errors. *Journal of Computational and Graphical Statistics.* 2007;16(2):265-288.

5. Crainiceanu CM, Goldsmith AJ. Bayesian Functional Data Analysis using WInBUGS. *Journal of Statistical Software.* 2010;32(11):1-33.

6. R-development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing (Vienna, Austria). 2009; ISBN 3-900051-07-0. (http://www.R-project.org). (Accessed January, 2010).

7. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for Running WInBUGS from R. *Journal of Statistical Software.* 2005;12(3):1-16.

Web Appendix 3.: Odds ratios (95% Confidence Interval) for peripheral arterial disease by urine cadmium quintiles using the 1/3 random sample with measured urine cadmium levels (N=2,098)

	Urine Cadmium, μg/g creatinine							
	Men				Women			
	Cases/ Non cases	Model 1	Model 2	Model 3	Cases / Non cases	Model 1	Model 2	Model 3
Quintile 1 (<0.20)	1 / 253	1 (ref)	1 (ref)	1 (ref)	8 / 89	1 (ref)	1 (ref)	1 (ref)
Quintile 2 (0.20 – 0.31)	12 / 221	4.17 (0.50, 34.54)	5.57 (0.66, 47.25)	5.21 (0.58, 47.01)	12 / 138	0.81 (0.24, 2.73)	0.68 (0.19, 2.40)	0.68 (0.18, 2.57)
Quintile 3 (0.31 – 0.44)	13 / 192	5.13 (0.58, 45.08)	5.08 (0.56, 46.14)	4.11 (0.49, 34.68)	14 / 211	0.51 (0.13, 2.06)	0.42 (0.09, 1.95)	0.38 (0.08, 1.81)
Quintile 4 (0.44 – 0.69)	15 / 207	6.68 (0.77, 57.68)	7.79 (0.82, 73.66)	5.99 (0.66, 54.19)	14 / 216	0.37 (0.09, 1.49)	0.31 (0.07, 1.41)	0.25 (0.05, 1.20)
Quintile 5 (≥ 0.69)	38 / 146	22.43 (2.29, 219.35)	22.92 (2.15, 244.20)	15.77 (1.57, 158.45)	29 / 269	0.74 (0.22, 2.54)	0.64 (0.14, 2.85)	0.40 (0.08, 1.98)
p-value linear trend*		< 0.001	< 0.001	0.003		0.98	0.92	0.34

* *P*-value for linear risk trend across quartiles of urine cadmium.
Model 1: Adjusted for age (years using restricted cubic splines with 5 knots), and race-ethnicity (white, black, Mexican-American, other).
Model 2: Further adjusted for education (< high school, ≥ high school), post-menopausal status for women (yes, no), body mass index (kg/m^2), blood lead, C-reactive protein (log-transformed), total cholesterol (mg/dl), HDL cholesterol (mg/dl), cholesterol lowering medication use (yes, no), systolic blood pressure (mm Hg), blood pressure lowering medication use (yes, no), diabetes (yes, no), and glomerular filtration rate (ml/min/1.73m^2).
Model 3: Further adjusted for smoking (never, former, current), and serum cotinine (log-transformed).