Education Corner

Reflection on modern methods: good practices for applied statistical learning in epidemiology

Yanelli Nunez,[1†] Elizabeth A Gibson,[1†] Eva M Tanner,[2] Chris Gennings,[2] Brent A Coull,[3] Jeff Goldsmith[4] and Marianthi-Anna Kioumourtzoglou[1]*

[1]Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, New York, NY, USA, [2]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA and [4]Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY, USA

*Corresponding author. Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, 722 W. 168th Street, New York, NY 10032, USA. E-mail: mk3961@cumc.columbia.edu
[†]Joint first authors.

Abstract

Statistical learning includes methods that extract knowledge from complex data. Statistical learning methods beyond generalized linear models, such as shrinkage methods or kernel smoothing methods, are being increasingly implemented in public health research and epidemiology because they can perform better in instances with complex or high-dimensional data—settings in which traditional statistical methods fail. These novel methods, however, often include random sampling which may induce variability in results. Best practices in data science can help to ensure robustness. As a case study, we included four statistical learning models that have been applied previously to analyze the relationship between environmental mixtures and health outcomes. We ran each model across 100 initializing values for random number generation, or 'seeds', and assessed variability in resulting estimation and inference. All methods exhibited some seed-dependent variability in results. The degree of variability differed across methods and exposure of interest. Any statistical learning method reliant on a random seed will exhibit some degree of seed sensitivity. We recommend that researchers repeat their analysis with various seeds as a sensitivity analysis when implementing these methods to enhance interpretability and robustness of results.

Key words: Statistical learning, random seed, environmental mixtures, machine learning, Bayesian statistics, penalized regression

Key Messages

- Statistical learning is increasingly useful for epidemiology applications as data dimensionality and complexity increase.
- Most statistical learning approaches incorporate random sampling. Defining a seed enables reproducibility.
- Findings may vary across seeds to different degrees, depending on the dataset and the chosen method.
- Sensitivity analyses should assess robustness of results to seed selection.
- If results are highly variable across seeds, a distribution of estimated effects across seeds should be presented.

Introduction

As the data we use for epidemiologic studies become more complicated—with high-dimensional exposure and outcome spaces and increasing sample sizes—investigators are reaching for statistical learning tools, i.e. a set of tools for modelling and understanding complex datasets more suited to accommodating big data.[1] Indeed, epidemiologists are increasingly adopting statistical learning methods to help answer research questions in public health. For example, epidemiologists have used clustering algorithms to determine the effect of atmospheric particulate matter with a diameter ≤ 2.5 micrometers ($PM_{2.5}$) exposure across cities with different $PM_{2.5}$ composition,[2] penalized regression to identify chemical-specific independent associations between environmental contaminants and birth weight,[3] and tree-based methods to assess potential interactions among air pollution toxins and their relationship with child cognitive skills,[4] among others. These approaches appear better equipped than traditional methods to accommodate numerous issues, such as model uncertainty, multicollinearity and multiple comparisons.

Most well-established methods for statistical learning incorporate random sampling, e.g. a tuning parameter that depends on a random division of the dataset or a random starting point to begin sampling.[5,6] When applying statistical learning to public health questions, understanding the role of random sampling in these methods and assessing the potential variability induced by this randomness can greatly increase confidence in the results.

Random dataset divisions are used by some statistical learning methods to split the full dataset to a 'training' dataset, to build the model, and a 'testing' dataset to (i) measure the predictive ability of said model and/or (ii) assist in estimating tuning parameters to maximize predictive accuracy. When no external testing dataset is available, cross-validation can be used in the full dataset; cross-validation randomly splits the original dataset in two, builds the model on one set of the data (i.e. the training set) and evaluates its performance on the left-out set (i.e. the testing set). A common extension, k-fold

cross-validation, randomly partitions the original data into k subsamples and repeats the training k times using all but a single subsample, while testing the model on the held-out sample. The predicted values for each held-out sample are then compared with the observed ones from the same sample, for example by averaging the k error estimates.[6] Cross-validation works well when the objective of the statistical analysis is prediction; it may not perform as well to assess sensitivity and specificity or to evaluate effect estimates. Nonetheless, cross-validation is often used for assessing goodness of fit and selecting tuning parameter values in epidemiologic models when there is a lack of other options to optimize statistical learning models for effect estimation.

The purpose of assessing model performance in a separate ('testing') dataset is to avoid over-fitting and ensure generalizability of results. In cross-validation, if each subsample is randomly drawn from the original dataset, then, on average, each should serve as a representative subsample. The key phrase here is 'on average'—it does not mean that any given subsample is representative of the complete sample. Researchers can enhance their confidence in the results simply by running a sensitivity analysis with a different seed, as the seed will determine the splitting. Similar results on tuning parameters chosen from different splits that arise from different seeds will increase confidence that the randomly chosen subsamples are representative of the whole and will strengthen conclusions regarding generalizability.

Statistical learning methods that rely on a Bayesian framework for statistical inference include—but are not limited to—Bayesian model averaging, e.g. to identify time windows of exposure to pollutants that produce adverse health effects;[7] Bayesian networks to account for uncertainty, as e.g. in infectious disease epidemiology;[8] and Bayesian hierarchical models, e.g. to investigate day-to-day changes in coarse particulate matter air pollution and cause-specific mortality.[9] These Bayesian methods involve a set seed for a different purpose than splitting the dataset into subsamples. Bayesian methods specify prior

distributions that, when combined with a data likelihood, give rise to a posterior distribution of interest. Calculating the exact posterior distribution of most models is impossible or, at least, too computationally expensive to be reasonable. Instead, Markov Chain Monte Carlo (MCMC) is used to produce samples from a model's posterior distribution. The marginal distribution of a Markov chain that has converged to the stationary distribution can be used to make inferences that approximate results from the full posterior distribution. MCMC is inherently random and chains often begin from a random point, which is defined by the set seed. In theory, if the chain runs long enough to converge, then the initialization seed should not matter.[5] In practice, convergence assessment is not always straightforward. Although there are diagnostic tools that quantify the variation among multiple chains to assess whether the chains have converged, such as the Gelman–Rubin statistic,[5,10,11] these are not always applied in epidemiologic studies. Similar results from a sensitivity analysis based on an initialization point assigned by a different seed will also increase confidence that the results represent the full posterior.

Assessing robustness of results from statistical learning methods in epidemiology can be improved by sensitivity analyses. Here, we present a case study of analytic examples of sensitivity to seed selection across four statistical learning models used in environmental epidemiology to assess exposure to environmental mixtures in health analyses, noting that this is but one example in public health. We discuss best practices to ensure robustness of results and the benefits of sensitivity analyses when utilizing statistical learning to make inference in public health research.

Methods

The aim of this analysis is to highlight the importance of incorporating seed sensitivity analyses into epidemiologic studies that employ statistical learning methods, using as a case study an environmental epidemiology application. We employed four methods that depend on a seed, for cross-validation or as a random start for a sampling chain, to illustrate variability in results across different seeds. We availed ourselves of existing models from a study presenting multiple methods to assess exposure to multipollutant chemical mixtures in environmental epidemiology. Specifically, we used information on 18 persistent organic pollutants (POPs) and data on leukocyte telomere length (LTL) among 1 003 US adults in the National Health and Nutrition Examination Survey (NHANES, 2001–2002).[12] LTL is the outcome variable in this study. In summary, LTL refers to the length of the repetitive nucleotide sequence at the end of chromosomes that provides stability

and allows for complete DNA replication.[13,14] The study population, POPs mixture and LTL measurements have been described in more detail previously.[15–22] According to previous work by Mitro et al., we a priori divided the 18 POP exposure mixture into three groups: (i) eight non-dioxin-like polychlorinated biphenyls (PCBs), (ii) two non-ortho PCBs (PCBs 126 and 169), and (iii) POPs with moderate to high toxic equivalency factors (mono-ortho-substituted PCB 118, four dibenzo-furans and three chlorinated dibenzo-p-dioxins), here referred to as mPFD.[12,15]

We built four models according to Gibson et al.[12] including the same 18 POPs as predictors of interest, LTL as outcome, plus covariates. We included two penalized regression methods (lasso and group lasso) for variable selection that add a regularization term with an associated tuning parameter (often denoted using λ) to a regression model and chose the tuning parameter using 10-fold cross-validation to control over-fitting. Lasso constrains the fit of a regression model with respect to the sum of the absolute values of the coefficients[23] and group lasso constrains the fit of the model with respect to the sum of the absolute values of a priori defined groups.[24] The regularization parameter λ in each model controls the degree of shrinkage and is tuned to improve model fit and predictive capacity. In both of these methods, we only penalized the POP variables, but not the potential confounders included in the model.

A third model [weighted quantile sum regression (WQS)] includes a single training and a hold-out set (testing set) to assess generalizability. WQS creates an empirically-weighted index of chemicals and includes this index as the exposure term in a regression model.[25] Here, cross-validation is not used to tune the model; instead, the hold-out dataset is used to determine whether weights generalize from one random subsample to another. WQS also estimates a parameter that constrains each chemical weight to be between 0 and 1 and all weights to sum to 1. Important chemical components in the mixture are identified by comparing the weight for each component with a threshold parameter, τ, chosen a priori. Here we use $\tau = 1/p$, where p is the number of chemicals in the mixture, as has been previously suggested.[12,26] Because WQS is inherently one-directional, in that it tests only for mixture effects positively or negatively associated with a given outcome, we specified a positive unconstrained model.

The fourth model we included, Bayesian Kernel Machine Regression (BKMR), uses MCMC sampling (with a random seed for initialization) to estimate the posterior distributions of the model parameters by simulating realizations from these intractable posteriors.[27] BKMR estimates chemical-specific exposure–response functions, detects potential interactions among mixture members, and estimates

the overall mixture effect. Its variable selection feature also outputs posterior inclusion probabilities. Posterior inclusion probability values range between 0 and 1 and their magnitude indicates relative variable importance.[28] For this analysis, we performed hierarchical variable selection for the three pre-defined groups of congeners. This feature outputs both group and congener-specific posterior inclusion probabilities, i.e. the relative importance of a congener given that the group that contains that congener is important.

We ran lasso, group lasso, WQS and BKMR 100 times each, using a different seed each time. We measured variability in lasso and group lasso results using the proportion of non-zero beta coefficients, the median and the inter-quartile range of the beta coefficients for each of the POPs across seeds. We measured variability in WQS results using (i) the proportion of weights above our chosen threshold ($\tau = 1/18$), the median and inter-quantile range for each POP across seeds, and (ii) the pooled WQS index coefficient across seeds calculated using Rubin's rule.[29,30] We measured variability in BKMR results by visualizing the full exposure–response curves and using the median and inter-quantile-range of the posterior inclusion probabilities for each of the POPs across seeds. All analyses were conducted in R version 3.6.1,[31] and all code to recreate results and figures is available online at https://github.com/yanellinunez/Commentary-to-mixture-methods-paper.

Results

As expected, for all methods, we observed some variability in the results based on seed number.

Lasso and group lasso

Lasso models varied slightly across seeds, beginning with different λ values chosen as the optimal tuning parameter (Figure 1). Whereas there was little difference in predictive accuracy across seeds (total cross-validation error range = 0.041–0.042), the chosen λ value based on cross-validation affected the number of coefficients that were pushed to zero. The smallest λ value selected, chosen by two of the 100 seeds, retained 9 of the 18 congeners in the model. The largest λ value selected, chosen by 11 of the 100 seeds, retained 4 congeners in the model. Sixty percent of all chosen λ values retained 5 congeners. However, independent of seed, 9 out of the 18 congeners consistently had beta coefficients of zero (dioxin 1,2,3,4,6,7,8-hpcdd, furans 1,2,3,4,7,8-hxcdf and 1,2,3,6,7,8-hxcdf, and PCBs 74, 138, 153, 170, 187 and 194); and 4 out of the 18 congeners (PCB 99, 118, 126 and furan 2,3,4,7,8-pncdf) consistently had non-zero beta coefficients. Four of the 5 other
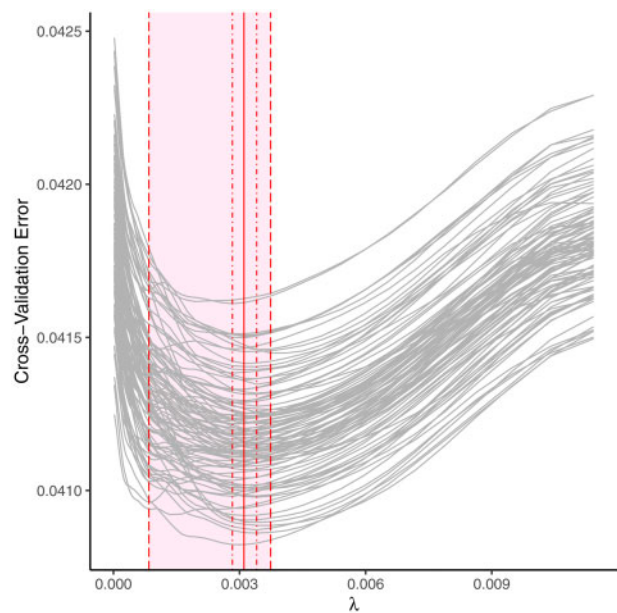

Figure 1 Cross-validation curves for lasso models over 100 seeds. Grey curves represent results obtained from each of the 100 seeds. The vertical red line indicates the median of the best fit λ values. The inner dot-dashed lines represent the inter-quartile range of the best fit λ values. The outer dashed lines represent the minimum and maximum of the best fit λ values. All best fit λ values fell inside the shaded area.

congeners, dioxins 1,2,3,4,6,7,8,9-ocdd and 1,2,3,6,7,8-hxcdd and PCBs 180 and 169 had non-zero coefficients in ≤10% of the cases. Only one congener, furan 1,2,3,4,6,7,8-hxcdf, had less consistency, with non-zero beta coefficients in 71% of the seeds (Figure 2).

In group lasso, the non-dioxin-like PCBs had the most variability across seeds relative to the other groups. This congener group had non-zero beta coefficients in 75% of the seeds. In contrast, the mPFD and non-ortho PCBs consistently had non-zero beta coefficients. The non-dioxin-like PCBs also showed the widest range of beta values for a given congener across seeds, particularly PCBs 180 and 153. The non-ortho PCBs had consistent coefficient values across seeds (Figure 2).

WQS regression

In WQS, weight values varied to a degree across seeds, which resulted in the number of weight values above the threshold τ and the congener order of importance varying from seed to seed. Out of the 100 seeds, furan 2,3,4,7–8-pncdf and PCB 126 had weight values above the threshold 96 and 68% of the time, respectively. These two congeners also had the largest weight in most instances—in 49 and 15 of the seeds, respectively. PCB 99, PCB 118 and furan 1,2,3,4,6,7,8-hxcdf had weights above the threshold in ~50% of the seeds and presented the largest weight value in 12, 8 and 4 instances, respectively. The remaining
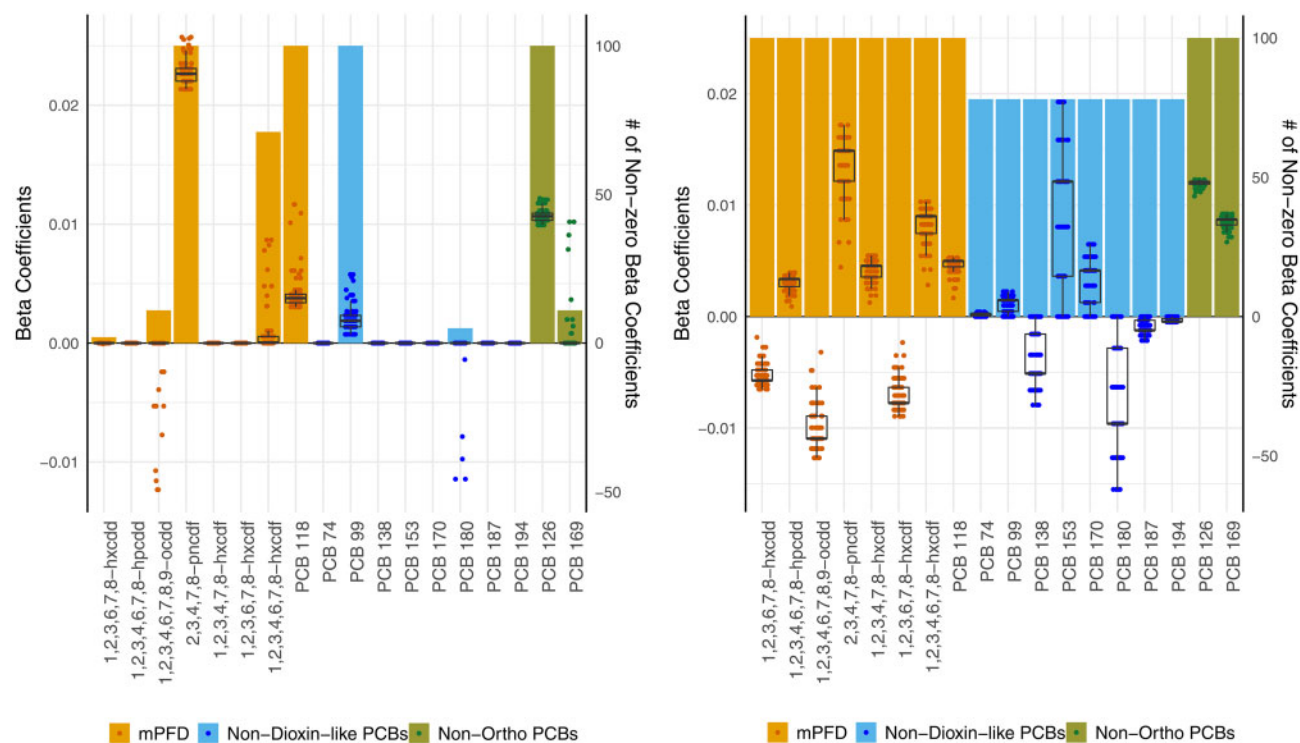
Figure 2 (**a**) Lasso and (**b**) group lasso beta coefficients over 100 seeds. Bars correspond to the right axis and indicate the number of times a congener had a non-zero beta coefficient across all seeds. Data points and boxplots correspond to the left axis. The data points are the congeners' beta coefficients in each of the seeds. Boxplots show the median and interquartile range for the beta coefficients of the given congener.

congeners showed weights above the threshold in <50% of the seeds (Figure 3). The regression coefficient for the pooled WQS index across seeds was significantly positive. The confidence intervals around the effect estimate for the WQS index did not include the null value for 97% of the seeds based on the hold-out data (Figure 4).

BKMR

For MCMC seeds, the order of posterior inclusion probabilities for the congener groups and the highest conditional posterior inclusion probability within each group (furan 2,3,4,7,8-pncdf for mPFD, PCB 126 for non-ortho PCBs and PCB 170 for non-dioxin-like PCBs) did not change based on seed selection (Table 1). Direction and magnitude of exposure–response functions for individual POPs across seeds were largely similar, with notable deviations for PCBs 99, 118, 126, 169 and 180, and furan 2,3,4,7,8-pncdf. For these 6 congeners, 4 of the 100 seeds produced null exposure–response curves with no suggestion of a positive or negative association with LTL. For all other seeds, the credible intervals around furan 2,3,4,7,8-pncdf did not include the null value at low levels of the exposure–response curve (Figure 5) and we observed suggestive evidence of associations with PCBs 126 and 169. For 96 of

100 seeds, we observed an overall mixture effect, with null results for the same four seeds.

Discussion

Our goal was to show the value of incorporating multiple seeds when applying statistical learning methods that involve a random process in epidemiologic studies. We used an environmental mixtures example as a case study and ran four methods over 100 different seeds, to illustrate the benefit of sensitivity analyses. Obtaining results across multiple seeds increases the generalizability of the results and the conclusions drawn. We showed that although setting a specific seed may ensure the reproducibility of an analysis, it does not guarantee generalizability and robustness, highlighting the benefit of estimating parameters for a given statistical learning model using multiple seeds.

For methods applying some form of cross-validation to select the optimal tuning parameter for the model, sensitivity to seed gives rise to variability in the choice of this parameter (Figure 1). Although there may be no practical significance in epidemiologic research between two λ values in terms of predictive accuracy measured by cross-validation error, two similar λ values may result in substantial differences in variable selection. In our case study,
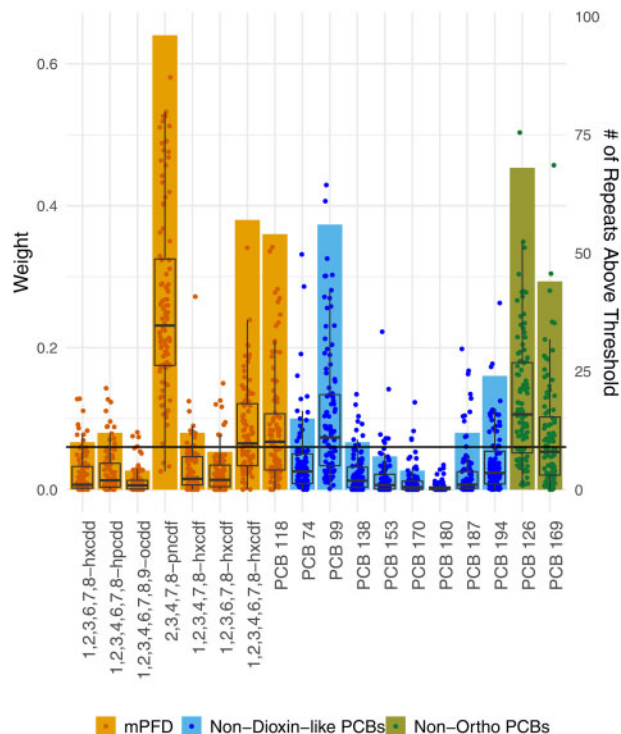
Figure 3 WQS estimated weights over 100 seeds. Bars correspond to the right axis and indicate the number of times a congener had a weight value above the threshold, which was calculated as $1/p$ ($1/18 = 0.05$, horizontal line) out of the 100 seeds. Data points and boxplots correspond to the left axis. The data points are the congeners' weights in each of the seeds. The boxplots show the median and interquartile range for the weights of each congener across seeds. WQS, Weighted Quantile Sum Regression.

the number of retained congeners varied between 4 and 9, depending on the optimal λ value selected by cross-validation error across seeds, with 5 congeners selected in 60% of the seeds. This could result in quite different conclusions about the potential toxicity of the examined congeners.

Variability in λ, thus, may affect the stability of beta coefficients across seeds (Figure 2). In the case of lasso, nine beta coefficients were consistently pushed to zero and four consistently pushed to non-zero, across all seeds. However, 1 congener (furan 1,2,3,6,7,8-hxcdf) had a positive beta coefficient for 71 out of 100 seeds and 4 other congeners (dioxins 1,2,3,4,6,7,8,9-ocdd and 1,2,3,6,7,8-hxcdd and PCBs 180 and 169) had non-zero coefficients <10% of the time. Thus, if this analysis were run with a single seed, it would have been possible to find and report that furan 1,2,3,6,7,8-hxcdf was not predictive of the outcome or that PCB 180 was. Depending on the context, the consequences of ruling out a potential association with one POP or drawing undue attention to another may be trivial or not. In the case of group lasso, all congeners within *a priori* defined groups were penalized together. Congeners in the mPFD and non-ortho PCB groups consistently had
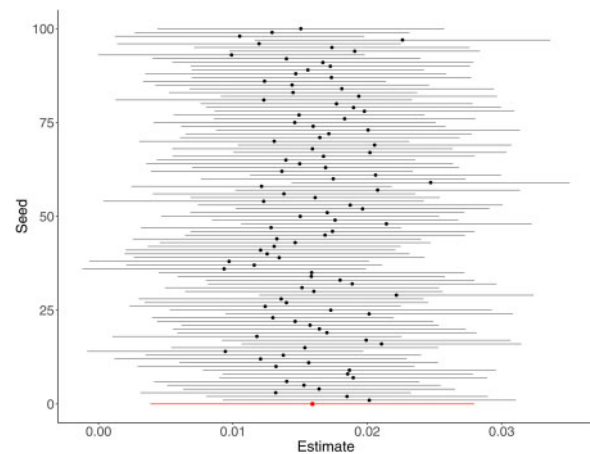

Figure 4 WQS index estimates over 100 different seeds. Black points and black lines represent estimates for each of the 100 seeds and the 95% confidence intervals, respectively. The red point represents the pooled estimate across all seeds and its 95% confidence interval. WQS, Weighted Quantile Sum Regression.

non-zero beta coefficients over the 100 seeds. Congeners in the non-dioxin-like PCBs group had non-zero beta coefficients in 75 seeds, but had beta coefficients of zero in the remaining 25 seeds. Again, if the analysis were run with only 1 seed, the predictive power of non-dioxin-like PCBs may have been missed.

In the case of WQS, the dataset is randomly partitioned into a training and a testing set. The weights for each congener are then calculated from the training set and subsequently used to estimate the index coefficient with the testing set. Thus, the results can be strongly influenced by the initial random partition of the data and may be unstable, especially when the dataset is small. Our analysis, with a sample size of 1003, showed that the number of congeners with a weight above the threshold and the magnitude of given weights varied across seeds. Tanner *et al.*[26,32] address this issue by using a repeated hold-out validation. This process repeats the analysis 100 times over randomly selected seeds, similar to what we did here, then reports the mean and confidence intervals for the sampling distribution of the weights instead of a single estimate from a single seed, enhancing generalizability. Code for this technique is also publicly available (https://github.com/evamtanner/Repeated_Holdout_WQS). Although we found some variability in the estimated congener-specific weights, the conclusion that this mixture had an overall harmful effect on LTL was robust to seed selection.

BKMR results varied little based on the selected seeds, but 4 of the 100 chosen MCMC seeds failed to identify any non-null univariate exposure–response functions or an overall mixture effect. Since seed should not influence results if the Markov chain has converged to its stationary distribution, we believe that the runs based on these 4 seeds

Table 1 Posterior inclusion probabilities for persistent organic pollutant groups and conditional posterior inclusion probabilities for individual congeners across 100 MCMC seeds

Group		Posterior inclusion probabilities		
		Minimum	Median	Maximum
mPFD[a]		0.65	0.86	0.89
Non-ortho PCBs[b]		0.61	0.67	0.76
Non-dioxin-like PCBs		0.41	0.46	0.68
Group	Congener	Conditional posterior inclusion probabilities		
		Minimum	Median	Maximum
mPFD	1,2,3,6,7,8-hxcdd	0.01	0.01	0.12
	1,2,3,4,6,7,8-hpcdd	0.01	0.01	0.13
	1,2,3,4,6,7,8,9-ocdd	0.004	0.01	0.12
	2,3,4,7,8-pncdf	0.12	0.86	0.89
	1,2,3,4,7,8-hxcdf	0.01	0.02	0.13
	1,2,3,6,7,8-hxcdf	0.01	0.02	0.13
	1,2,3,4,6,7,8-hxcdf	0.01	0.02	0.14
	PCB 118	0.05	0.06	0.14
Non-ortho PCBs	PCB 126	0.51	0.65	0.68
	PCB 169	0.32	0.35	0.49
Non-dioxin-like PCBs	PCB 74	0.08	0.10	0.13
	PCB 99	0.11	0.13	0.16
	PCB 138	0.10	0.12	0.14
	PCB 153	0.12	0.14	0.16
	PCB 170	0.12	0.17	0.21
	PCB 180	0.11	0.13	0.17
	PCB 187	0.08	0.09	0.13
	PCB 194	0.09	0.10	0.12

[a]Mono-ortho-substituted PCB 118, dibenzo-furans, and chlorinated dibenzo-p-dioxins
[b]Polychlorinated biphenyls

did not converge. It is likely that the chain became 'stuck' in an area with a local optimum and 100 000 iterations were not sufficient to examine the full distribution. Further inspection of convergence criteria, however, provided no indication of failure to converge, supporting our recommendation of sensitivity analyses with multiple seeds or the use of convergence diagnostics, such as the Gelman–Rubin statistic.

Recommendations of this sort exist already in various subject areas. In machine learning, re-fitting a model, such as a penalized regression or a tree-based method, on bootstrapped samples of the data has been recommended.[24,33] This is another means of capturing variability attributed here to seed selection. A single regression tree, for example, is not very stable; a different seed will divide the data into different training and testing sets that may result in vastly different trees. A random forest, which trains multiple trees on bootstrapped samples of observations and uses only a random subset of variables at each split, is much more robust. The method we used to pool WQS coefficients across

seeds, Rubin's rule, is commonly implemented in stochastic missing-data imputation to combine variance between and within multiple imputations.[29,30] For Bayesian MCMC, it is often recommended to simulate multiple chains and then check that they have converged to the same distribution, i.e. 'mixed'.[34] In this reflection on seed selection in statistical learning, we aim to connect applied epidemiologic research to these avenues.

A degree of variability across seeds is expected due to the intrinsic randomness associated with the methods at hand. However, seed sensitivity is not specific to these four methods; any statistical learning tool may be susceptible. Nor is the generalizability and robustness of results only method-dependent—factors such as sample size or data heterogeneity should also be considered. Finally, we note that many statistical learning methods, including lasso and group lasso, were developed to improve predictive accuracy. Their ability to accommodate complex and high-dimensional datasets make them increasingly appealing tools for use in epidemiologic analyses; application,
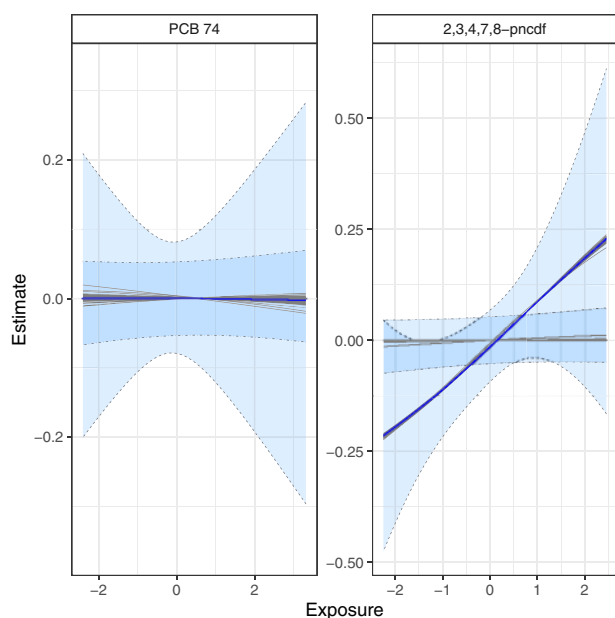
Figure 5 Congener-specific effect estimates for two mixture members, PCB 74 (non-dioxin-like PCBs) and furan 2,3,4,7,8-pncdf (mPFD) estimated by BKMR over 100 MCMC seeds. Grey lines represent exposure–response functions for each of the 100 seeds. The blue line indicates the median across all seeds. The area between the dashed lines includes the 95% credible intervals for 96 seeds where the credible intervals around furan 2,3,4,7,8-pncdf did not include the null value at low exposure levels. The narrower dot-dashed lines represent the 95% credible intervals for the four seeds that produced null exposure–response curves for furan 2,3,4,7,8-pncdf. BKMR, Bayesian Kernel Machine Regression; MCMC, Markov Chain Monte Carlo.

nonetheless, to estimate health effects should be performed with caution.

Conclusion

Running the same analysis under different seeds provides a better understanding of the results. Thus, we recommend that epidemiologists employing statistical learning methods run models that involve a random component with multiple seeds as best practice. Results across seeds should not be used to select a seed number, but instead as sensitivity analysis to assess the robustness of the results and enhance generalizability of study findings. A randomly selected seed should be used for the main analysis, and results across seeds should be included as supplementary material. When results across seeds vary greatly, researchers should consider reporting averages and inter-quartile ranges rather than an estimate from a single seed.

Funding

Conflict of interest

Nothing to declare.

References

1. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. vol. **112**. New York, NY: Springer, 2013.
2. Kioumourtzoglou MA, Schwartz J, James P, Dominici F, Zanobetti A. PM2.5 and mortality in 207 US cities: modification by temperature and city characteristics. *Epidemiology* 2016;**27**: 221.
3. Lenters V, Portengen L, Rignell-Hydbom A *et al*. Prenatal phthalate, perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: multi-pollutant models based on elastic net regression. *Environ Health Perspect* 2016;**124**:365–72.
4. Stingone JA, Pandey OP, Claudio L, Pandey G. Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. *Environ Pollution* 2017;**230**:730–40.
5. Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall/CRC; 2013.
6. Murphy KP, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
7. Dominici F, Wang C, Crainiceanu C, Parmigiani G. Model selection and health effect estimation in environmental epidemiology. *Epidemiology* 2008;**19**:558–60.
8. Lau CL, Smith CS. Bayesian networks in infectious disease eco-epidemiology. *Rev Environ Health* 2016;**31**:173–77.
9. Chen R, Yin P, Meng X *et al*. Associations between coarse particulate matter air pollution and cause-specific mortality: a nationwide analysis in 272 Chinese cities. *Environ Health Perspect* 2019;**127**:017008.
10. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statist Sci* 1992;**7**:457–72.
11. Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. *J Am Stat Assoc* 1996;**91**: 883–904.
12. Gibson EA, Nunez Y, Abuawad A *et al*. An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length. *Environ Health* 2019;**18**:76.
13. Blackburn EH. Telomere states and cell fates. *Nature* 2000;**408**: 53–6.
14. Greider CW. Telomere length regulation. *Annu Rev Biochem* 1996;**65**:337–65.
15. Mitro SD, Birnbaum LS, Needham BL, Zota AR. Cross-sectional associations between exposure to persistent organic pollutants and leukocyte telomere length among US adults in NHANES, 2001–2002. *Environ Health Perspect* 2016;**124**:651–58.
16. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J. Health and nutrition examination survey plan and operations, 1999-2010. *Vital Health Stat* 2013;**1**.
17. CDC. Atlanta NCfEH GA: CDC, editor. Laboratory Procedure Manual: PCBs and Persistent Pesticides in Serum. 2001—2002,

2002. https://www.cdc.gov/nchs/data/nhanes/nhanes_01_02/l28poc_b_met_pcb_pesticides.pdf (28 November 2018, date last accessed).

18. CDC. Atlanta NCfEH GA: CDC, editor. Laboratory Procedure Manual: PCDDs, PCDFs, and cPCBs in Serum. 2001—2002, 2002. https://www.cdc.gov/nchs/data/nhanes/nhanes_01_02/l28poc_b_met_dioxin_pcb.pdf (28 November 2018, date last accessed).

19. Akins JR, Waldrep K, Bernert JT. Jr The estimation of total serum lipids by a completely enzymatic 'summation' method. *Clin Chim Acta* 1989;**184**:219–26.

20. Cawthon RM. Telomere measurement by quantitative PCR. *Nucleic Acids Res* 2002;**30**:e47.

21. Lin J, Epel E, Cheon J *et al.* Analyses and comparisons of telomerase activity and telomere length in human T and B cells: insights for epidemiology of telomere maintenance. *J Immunol Methods* 2010;**352**:71–80.

22. Needham BL, Adler N, Gregorich S *et al.* Socioeconomic status, health behavior, and leukocyte telomere length in the National Health and Nutrition Examination Survey, 1999–2002. *Soc Sci Med* 2013;**85**:1–8.

23. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc Ser B (Methodol)* 1996;**58**:267–88.

24. Friedman J, Hastie T, Tibshirani R, *The Elements of Statistical Learning.* 2nd ed. Springer Series in Statistics New York, NY, USA, 2001.

25. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J Agric Biol Environ Statist* 2015;**20**:100–20.

26. Tanner EM, Bornehag CG, Gennings C. Repeated holdout validation for weighted quantile sum regression. *MethodsX* 2019;**6**:2855–60.

27. Bobb JF, Valeri L, Claus Henn B *et al.* Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* 2015;**16**:493–508.

28. Bobb JF, Henn BC, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ Health* 2018;**17**:67.

29. Rubin DB. *Multiple Imputation for Survey Nonresponse.* New York: Wiley, 1987.

30. Barnard J, Rubin DB. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999;**86**:948–55.

31. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2019. https://www.R-project.org/.

32. Tanner EM, Hallerback MU, Wikstrom S *et al.* Early prenatal exposure to suspected endocrine disruptor mixtures is associated with lower IQ at age seven. *Environ Int* 2020;**134**:105185.

33. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. *J Am Stat Assoc* 2011;**106**:608–25.

34. Levin DA, Peres Y. Markov chains and mixing times. *Am Math Soc* 2017;**107**.