Case study 1: Effect of chemical exposures on preterm birth

Olivier Binette, Brian Kundinger and Joe Mathews

abstract...

1. Introduction

This study attempts to assess the effect of dichlorodiphenyldichloroethene (DDE) and polychlorinated biphenyls (PCBs) on the risk of preterm birth for pregnant women. This paper will first provide findings from exploratory data analysis, discuss challenges for our analysis given these findings. We then conduct logistic regression on the data provided, assess the validity of those results through Bayesian Model Averaging, and then compare these results to those found through a Random Forest Model.

2. Materials and Methods

2.1 Data

Our analysis is based on the *Longnecker* dataset (Longnecker et al., 2001), which contains information on a subset of 2380 pregnant women enrolled in the National Collaborative Perinatal Project (CCP). It relates gestational age to chemical exposures (*dde* and *pcb_** concentrations), to socio-economic variables (*maternal age*, *race*, *smoking status*, *center* of enrollment, and education, income and occupation *score_** variables), as well as to health-related variables (*cholesterol* and *triglycerides* concentrations).

Cleaning up the data, we removed the *albumin* variable, which contained 93% missing values, and dropped observations of pregnancy over 55 week. The single case with missing *pcb_** values was also removed. We are left with $n = 2374$ observations and only missing values in the *score_** variables (22%). *Preterm* birth is defined as gestational age strictly less than 37 weeks ($n = 361$; 15%), and *very preterm* is defined as gestational age strictly less than 34 weeks ($n = 103$; 4%).

To facilitate data visualization and interpretation of the results, we grouped together the "black" ($n = 1220$) and "other" ($n = 123$) race categories into the single "non-white" category. For similar reasons, we summed together the 11 positively correlated *pcb_** variables, obtaining the *totalpcb* variable. While this preserves units, important information may be lost through this process.

Challenges and limitations. Our scientific understanding of this data and of the data collection process is limited by a lack of documentation. It not possible to identify the represented population without knowledge of the enrollment mechanism, and therefore our observations may not generalize. Futhermore, it is unclear what should be taken as a scientifically meaningful predictor which strongly correlates to environmental exposures. Should we directly consider *dde* or should it be normalized by lipid concentration? This is not something we can address through the data. Given these limitations and the fact that we have explored many models with the stated goal of identifying an interaction between *dde* and preterm birth risk, the results of our analysis should be taken as highly tentative and exploratory in nature.

Futhermore, there is a noticeable partial correlation between the *dde* and *totalpcb* variables ($\rho = 0.3$), after controlling for the linear effect of the other predictors (excluding *gestational_age*). This could be the shadow of an unobserved confounding variable: unless *dde* and the *pcb_** are breakdown products of the same exposure, this hints at another factor contributing to higher chemical concentrations in the blood. This prevents us from singling out any potentially causal effect.

2.2 Missing Values Imputation

The 22% missing values in the *score_** variables were imputed under a missing at random assumption using a standard Bayesian framework. We treated the score variables as independent, identically distributed

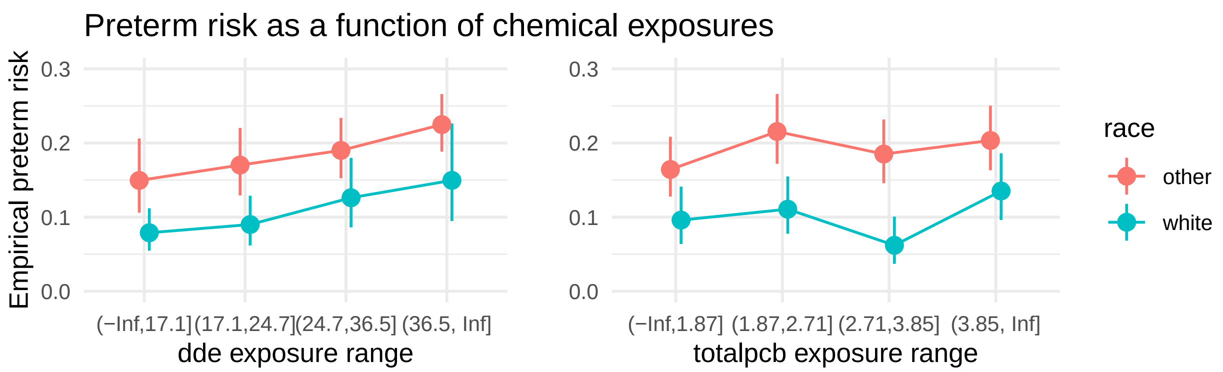


Figure 1: Marginal relationship between empirical preterm risk and chemical exposures, without controlling for socio-economic and health variables. The exposure ranges have been defined so that each class contains 25% of the observations and vertical lines represent 95% confidence intervals for the estimated risk.

Normal random variables and fitted a Bayesian linear model using the other predictor variables and a non-informative prior. That is, we regressed the observed values of the *score_** variables onto the other predictor variables and treated missing score values as model parameters, estimated through their posterior mean.

This approach has two major limitations. First, separating imputation from model fitting prevents the propagation of imputation uncertainty, although the Bayesian formulation would allow to combine the two steps together. Second, our approach assumes that observed variables in the data set can adequately predict the missing score values, or that data is missing at random. Even if this were to be the case, the data does not suggest clear linear relationships between the *score_** variables and other predictors. The potential of more complex non-linear models for the analysis of this dataset is considered in the Discussion section.

2.3 Logistic regression model for preterm outcome.

We model the log odds ratio of preterm birth as a linear function of the other covariates and perform maximum likelihood estimation. This allows us to control for the baseline effect of the *race*, *smoking_status* and *center* factors, and to control for the linear effect of socio-economic and health-related variables (*maternal_-age*, *score_income*, *score_education*, *score_occupation*, *cholesterol*, and *triglycerides*) on the log odds ratio. Appart from *dde* and *totalpcb*, no other variables are incorporated into this base model, and no interaction terms are considered.

The importance of the *dde* and *totalpcb* variables is assessed by: (1) testing for individual and joint statistical significance of the variables; (2) by transforming the associated p-values to the more meaningful $B(p) = -ep\log(p)$ scale; and (3) by interpreting the estimated effect size. For part (1), individual statistical significance is based on standard approximate t-tests and joint significance is tested with an approximate χ^2 test. The rationale for (2) is that $B(p)$ provides a lower bound on the Bayes factor comparing the null hypothesis of no effect to the alternative (over a large nonparametric class of reasonable prior distributions and when $p < e^{-1}$; see Sellke et al. [2001]). For example, a p-value of 0.01 is transformed to $B(0.01) \approx 1/8$: if both the null and the alternative hypotheses are a priori equally likely, then a posteriori the alternative hypothesis cannot be more than 8 times more likely than the null. Our interpretation (3) relies on the multiplicative contribution of the *dde* and *pcb* coefficients on the odds ratio of preterm birth, as a function of the empirical quantiles of exposure.

3. Results

Figure 2 below shows the clinically significant estimated effects of the *dde* and *totalpcb* variables. The estimated odds ratio, i.e. the estimated probability of preterm birth to non-preterm birth, can more than double as exposure to *dde* or *totalpcb* goes from zero to some of the larger concentration values observed in the dataset. This is roughly comparable to the marginal effect (which does not take into account the control variables) shown in the left panel of Figure 1. Note that the *dde* and *totalpcb* effect estimators are correlated ($\hat{\rho} = -0.29$), contributing to the rather large width of the confidence intervals. Furthermore, we warn that the estimated multiplicative effect in the tail may be unreliable. While we expect the logistic regression to properly capture overall trends, there is not reason for it to adequately fit more particular features of the data.

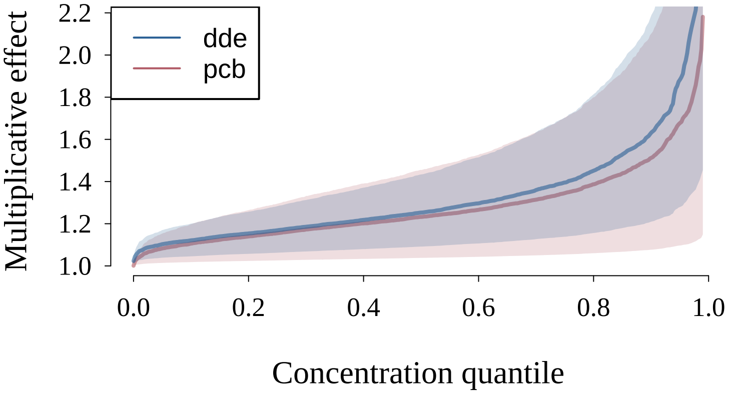


Figure 2: Multiplicative effect of dde and totalpcb on the odds ratio of preterm to non-preterm birth, as a function of the empirical dde or pcb concentration quantile. Confidence intervals are represented by shaded regions, and the plot is truncated on the right at the 0.99 quantile.

Furthermore, the *dde* and *totalpcb* variables have statistically significant effects: it is unlikely that a trend of this order would have arised by chance under the logistic model. The p-values computed can be interpreted on the $-ep\log(p)$ scale: the Bayes factor in favor of a non-null effect could be up to 47 to 1 for the *dde* variable, and up to 5.8 to 1 for the *totalpcb* variable.

	dde	*totalpcb*	jointly
p-value	$1.2 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$2.27 \cdot 10^{-5}$
$B(p)$	$1/47$	$1/5.8$	$1/1514$

The effect of *dde* and *totalpcb* on preterm risk found in this data, after superficially controlling for confounding factors, is very concerning. However, we cannot currently single out any of these two variables as having a distinct effect on preterm birth. The results should also be considered with care, as we do not expect the logistic model to correctly capture the data-generating mechanism. Inference about the “existence” of an effect is not meaningful in this context, and our analysis should be understood as only showcasing a particular aspect of this data.

4. Discussion

We used a logistic regression to estimate the effect of *dde* and *totalpcb* on preterm risk, controlling for socio-economic, health-related variables, and center heterogeneity (through their baseline effect). This allowed us to display a significant association between presence of these chemicals and preterm risk: higher concentrations increased the estimated risk of preterm birth. A further analysis, based on a hierarchical model, would also allow model coefficients for control variables to vary across centers.

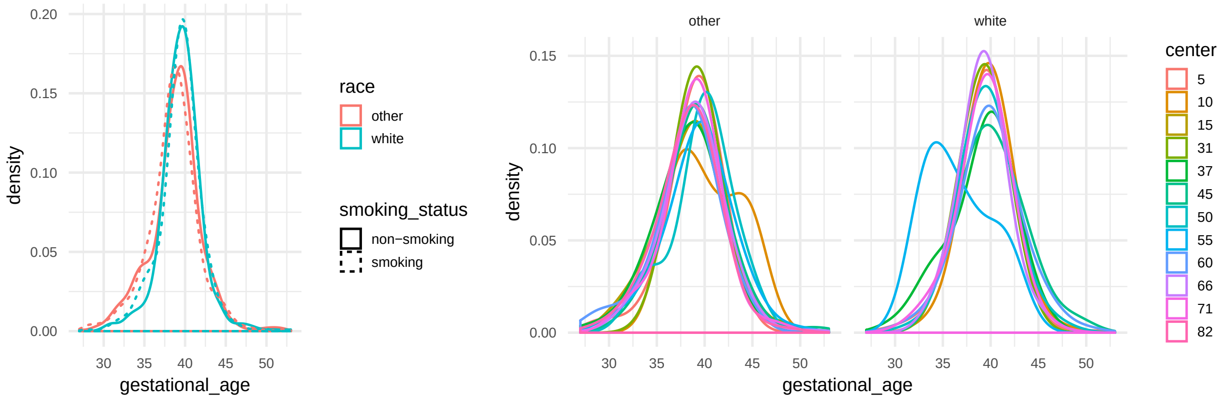
However, important limitations prevent us from drawing definitive conclusions from this analysis. In addition to the issues previously mentioned, our methodological approach could have been improved. First, it is clear that the logistic model, without any transformation of variables or interactions, does not correspond to a plausible data-generating model. In this context, the p-values and Bayes factors computed conditionally on null logistic models are not very meaningful, and a non-parametric bootstrap approach to the computation of p-values and confidence intervals would have provided more adequate quantification of uncertainty. Second, we could consider non-linear models to better control for the socio-economic and health-related variables. This is considered in the appendix with a random forest regression of gestational age, and a heuristic is proposed to assess variable significance in this context.

Appendix

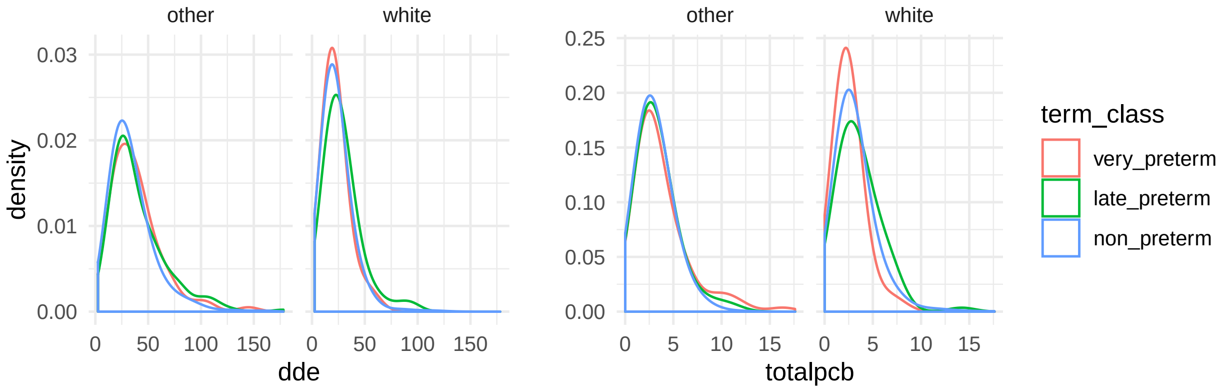
Data visualization

We plot some aspect of the data below, showcasing the marginal association between some of the variables and gestational age.

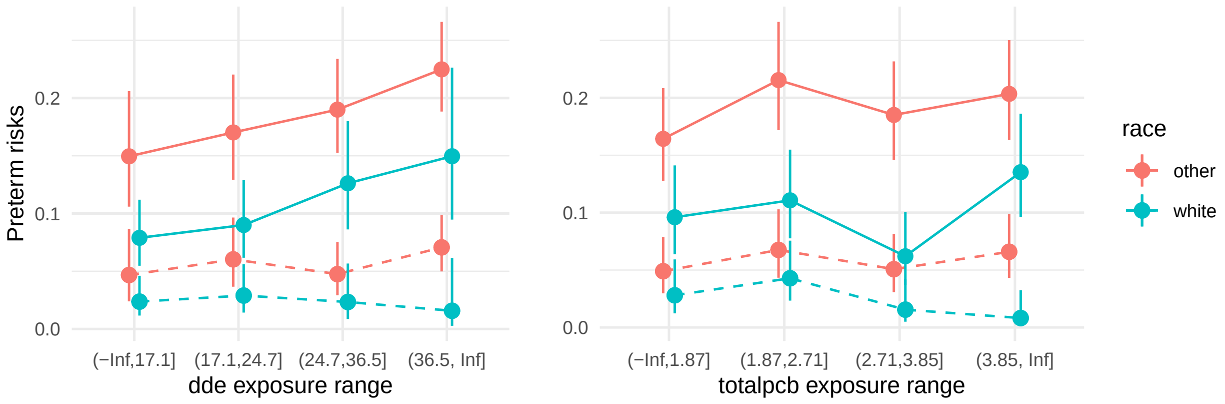
First, we consider the gestational age distribution disaggregated by race. The left panel below further distinguishes smoking and non-smoking women, and the right panel shows the gestational age distribution in the different centers. We notice that non-white women have a much heavier left tail of gestational age, corresponding to higher preterm risk. Smoking also seems to affect preterm risk, but to a lesser extent. On the right, we notice slight heterogeneity between centers, with only a few of them notably differing from the others (centers 10 and 55).



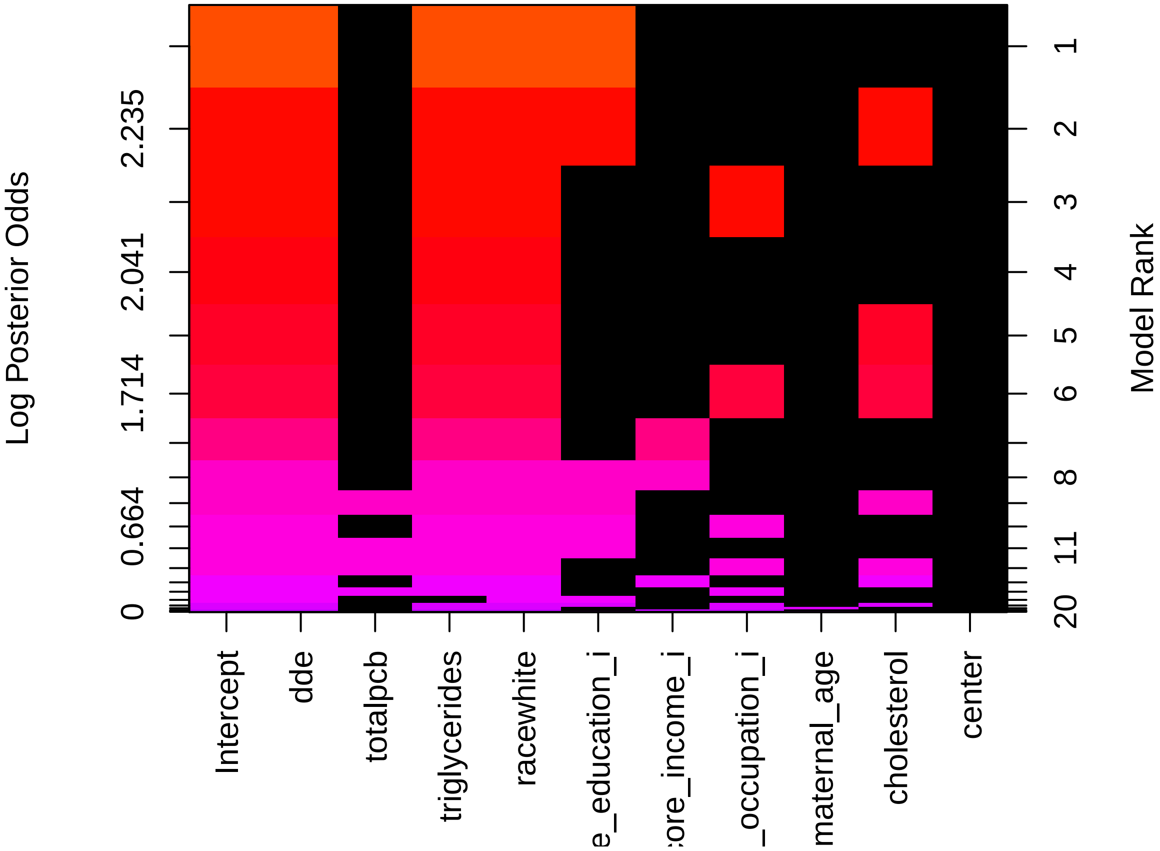
Next, we plot the distribution of chemical concentrations, disaggregating among women with late preterm ($\geq$ 34 weeks and $<$ 37 weeks), very preterm ($<$ 34 weeks) or non-preterm births. For women with late preterm birth, the chemical concentration distribution seems heavier on the right. This is not clearly seen for women with very preterm births, although not seeing an effect here may be due to the low sample size in this category.



Finally, we reproduce below Figure 1, which shows preterm risk as a function of range of chemical concentrations, adding to it the empirical very preterm risks. There is no clear marginal relationship between chemical concentrations and very preterm risks.



Bayesian model averaging



	dde	*totalpcb*	*race*
Inclusion Probability	0.96	0.20	0.98

Since cases classified as preterm and early preterm comprise relatively small porportions of the data, and since we reduced the individual PCB variables into one aggregate totalpcb variable, we check the significance of these variables through Bayesian Model Averaging (BMA). This process works by assigning a prior distribution to the set of possible models, and then uses the observed data set to calculate posterior probabilities for each of the possible models. One advantage of this method is that we obtain a posterior probability for each variable, which serves as a way to quantify undercertainty about each variable's statistical significance.

Using a uniform prior distribution on the set of models, we conduct BMA on the models outlined above. The posterior probabilities of each variable are provided below.

BMA mostly confirms the statstical significances found through logistic regression, but notably highlights uncertainty about the explanatory power of PCB in the model. We believe this inconsistency is due to the low number of preterm and earlypreterm births observed in the dataset, and note that further study should be done to asses the statistical significance of PCB.

In comparing results based on different definitions of preterm birth, we find that it is likely the case that DDE (and to a lesser extent, PCB) are associated with early birth. However, in the more extreme cases, where gestational age is less than 34 years, there are likely other factors not represented in the data that account for this difference.

Non-linear interactions

We use a random forest to regress *gestational_age* onto the other variables (the same used as in the main analysis). In comparison to the abysmall predictive accuracy of the logistic regression model (85%, which is barely better than a trivial classifier in this context), the random forest model correctly classifies 95.5% of the cases over the training data. While there could be overffiting here, this is inconsequential in the interpretation of the following analysis.

In order to assess the importance of the dde variable in this context, we consider the distribution of the random forest predictive accuracy when *dde* is replaced by white noise. This is plotted below.

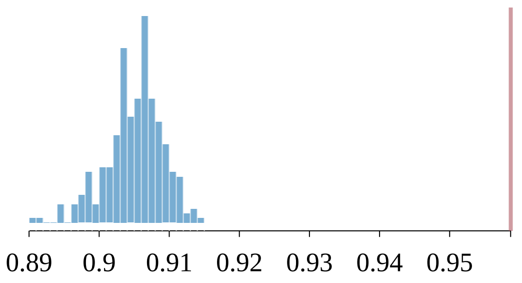


Figure 3: Distribution of the random forest predictive accuracy when dde has been replaced by white noise. The predictive accuracy of the random forest model with the dde variable is marked by a red vertical line.

This shows *dde* has a much more significant effect on predictive accuracy than a random noise covariate would have, even after controlling for the other confounding variables. The results obtained in the main analysis are therefore somewhat robust to the consideration of non-linear relationships in the data.

The above heuristic is not exactly standard, but it can be interpreted, for example, in the context of a standard linear model with gaussian errors. Using the R^2 goodness of fit statistic in this linear model context, the above procedure is entirely equivalent to a standard F-test. For non-linear models, it is not quite as clear wether a white noise null hypothesis is appropriate. Randomly permuting the dde values could alternatively provide a distribution for the random noise covariate. This doesn’t noticeably change the results in our case.

References

Thomas Sellke, M. J Bayarri, and James O Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.