Variable selection for discrete survival model with frailty in presence of left truncation and right censoring: Studying association of environmental toxicants on time-to-pregnancy

Abhisek Saha | Rajeshwari Sundaram

Understanding the association between mixtures of environmental toxicants and time-to-pregnancy (TTP) is an important scientific question as sufficient evidence has emerged about the impact of individual toxicants on reproductive health and that individuals are exposed to a whole host of toxicants rather than an individual toxicant. Assessing mixtures of chemical effects on TTP poses significant statistical challenges, namely (i) TTP being a discrete survival outcome, typically subject to left truncation and right censoring, (ii) chemical exposures being strongly correlated, (iii) appropriate transformation to account for some lipid-binding chemicals, (iv) non-linear effects of some chemicals, and (v) high percentage of concentration below the limit of detection (LOD) for some chemicals. We propose a discrete frailty modeling framework (named Discnet) that allows selection of correlated covariates while appropriately addressing the methodological issues mentioned above. Discnet is shown to have better and stable false negative and false positive rates compared to alternative methods in various simulation settings. We did a detailed analysis of the pre-conception endocrine disrupting chemicals and TTP from the LIFE study and found that older females, female exposure to cotinine (smoking), DDT conferred a delay in getting pregnant, which was consistent across various approaches to account for LOD as well as non-linear associations.

KEYWORDS
chemical exposure, discrete survival, elastic net, frailty, limit of detection, time-to-pregnancy

1 | INTRODUCTION

Synthetic environmental persistent chemicals, such as organochlorine pesticides, polychlorinated biphenyls (PCBs) have been reported to have endocrine-disrupting properties according to the Endocrine Society.1,2 Growing evidence indicates exposures to endocrine-disrupting chemicals (EDCs) may adversely affect human health such as reproduction, metabolism to name a few. EDC exposure is widespread as the result of environmental distribution through natural processes, bio-accumulation long after halted industrial production, as well as continual production of commonly used materials and goods containing a variety of EDCs contributes to environmental ubiquity.3 Data from human...
biomonitoring studies suggest women of childbearing age are often exposed to multiple EDCs simultaneously.\textsuperscript{4,5} Thus, this necessitates detailed study quantifying the risk of adverse reproductive outcomes in the context of EDC mixtures. Motivated by these issues, we are interested in studying the association of mixtures of EDCs with human fecundity, the biologic capacity to reproduce by men and women.\textsuperscript{6}

Assessing EDC mixtures association with human fecundity presents multiple statistical challenges. Human fecundity is quantitatively assessed through time-to-pregnancy (TTP), the number of menstrual cycles needed to get pregnant by non-contraceptive couples. This is a count type of data, further subject to censoring due to loss to follow-up or end of study and left truncation due to delayed entry into the study, leading to a left truncated, right censored discrete survival time. Thus, the statistical models and inference procedures need to account for this inherently discrete survival time. Additionally, it is well known that human fecundity has significant unmeasured heterogeneity.\textsuperscript{7} This necessitates including frailty in the model for TTP. While EDCs are highly correlated within an individual, as well as with the partners’ exposure levels, they further present quite a many challenges when it comes to modeling risk, such as (i) high concentration of values below limits of detection (LOD) in ascertaining the EDC levels: machines reported values may not be reliable, (ii) lipophilic nature of some EDCs, that is, they bind to the lipids in the blood: high amount of lipid will induce high amount of chemicals, thus needing to appropriately transform chemical levels as a function of lipid (iii) suspected non-linear effects of some EDCs, and (iv) missing covariates. As modern studies are capable of collecting large number of EDCs, modeling risk in the context of studying EDC mixtures requires a variable selection strategy which can handle the nuances of such data.

To bridge this gap, we propose a variable selection approach, called discnet to assess the mixtures of EDCs on TTP, which allows us to identify the “important drivers” of the mixtures of EDCs on human fecundity.\textsuperscript{8} To the best of our knowledge such analysis has not been performed before due to absence of variable selection methods addressing the mentioned issues. In this context, our proposed method is innovative leading to novel findings, that are consistent to prior research done in low dimensional variables settings.

We first note that such a problem can be converted to a penalization regression framework with modified generalized linear mixed model prior to adjusting for LOD and lipid binding phenomena. Although, there have been significant literature on variable selection using penalized regression framework spanning over few decades (in linear model,\textsuperscript{9,10} in generalized linear model,\textsuperscript{11-13} in continuous Cox proportional hazard model\textsuperscript{14,15}), relatively far less focus was given to variable selection in the context of survival model in presence of frailty. In this context, Fan and Li\textsuperscript{16}’s proposed method based on non-concave penalty is particularly noteworthy. Although it incorporates frailty in the penalized framework it is not applicable to discrete survival time as it addresses Cox’s proportional hazard model for continuous survival time. In the same context, Groll and Tutz’s work is particularly relevant as they developed L1-penalty based variable selection for generalized linear mixed models (GLMMs).\textsuperscript{17} They later adapted it for variable selection for discrete survival model with frailty,\textsuperscript{18} henceforth referred to as glmmlasso. L1-based method usually performs poorly in presence of highly correlated variables,\textsuperscript{10} and they do not produce stable paths.\textsuperscript{11} In our framework Discnet, we propose fast variable selection method based on elastic net penalty that can incorporate left truncation in addition to right censoring while addressing the concerns such as LOD and lipid treatment by incorporating cutting-edge procedures on the covariates. Our methodology also differs from that of glmmlasso in how the tuning parameters are chosen. We have used permutation selection method\textsuperscript{19} as it favors moderate sized models and a more balanced trade-off between power and false discovery rate. We have provided with the key R-function, Discnet.R with necessary library support of auxiliary R-functions, which can be downloaded from this github link (https://github.com/sahaAbh/Discnet).

In Section 2, we present our data structure and proposed penalized framework. In Section 3, we present necessary estimation framework that discusses computational schemes and, finally extend it to group selection. In Section 4, we present detailed finite sample investigation of our proposed approach through extensive simulations, and evaluate its strengths compared to the competing method. In Section 4, we have thoroughly investigated our proposed variable selection approach in the context of mixtures of PCBs from both partners and their TTP from The Longitudinal Investigation of Fertility and the Environment (LIFE) study.\textsuperscript{20} We have systematically addressed the issues of LOD, lipid adjustment to account for lipophilic EDCs and issues for non-linearity, as well as account for missingness through multiple imputation. Finally, in Section 6, we summarize our novel contributions, and key findings.

2 | METHOD

Let $T$ be the discrete time to event (eg, number of menstrual cycles needed to get pregnant), taking values in $\{1, \ldots, k\}$. let $\mathbf{x}$ be the vector of underlying covariates. In discrete survival analysis, one commonly models the discrete hazard function,
\(\lambda(t|\mathbf{x})\), and estimates the survivor function \(S(t)\),

\[
\hat{\lambda}(t|\mathbf{x}) = P(T = t| T \geq t, \mathbf{x}),
\]

\[
S(t|\mathbf{x}) = P(T \geq t|\mathbf{x}) = \prod_{i=1}^{l-1}(1 - \lambda(t|\mathbf{x})), \quad t = 1, \ldots, k.
\]

We consider here the log odds model for the hazard, which is the Cox model for the discrete survival as given below,\(^{21}\)

\[
\hat{\lambda}(t|\mathbf{x}) = h(\gamma_t + \mathbf{x}^T \beta), \quad h = g^{-1}, \quad g = \text{logit link function},
\]

where \(\gamma_t\) is unspecified baseline hazard, and \(\beta\) is the vector of coefficient corresponding to \(\mathbf{x}\). We further assume \(T\) is subject to both right censoring and left truncation. Let \(L\) and \(C\) be the non-informative left truncation random variable and non-informative right random censoring variables.\(^{22}\) Note, in presence of left truncation \(T\) is observed only if \(T \geq L\). Due to right censoring, \(\min(T, C)\) is only observed. We further assume that \((L, C)\) is independent of \(T\) and that \(P(L \leq C) = 1\). Let \(\delta\) be the censoring indicator, \(\delta = I(T \leq C)\) and \(\hat{T} = \min(T, C)\) be the observed time to event variable. Thus, \(\delta = 1\) implies \(L \leq T \leq C\) while \(\delta = 0\) implies \(T > C\). When \(T < L\) the sample is not observed. Let the \(i\)th sample be represented by the triplet \((t_i, l_i, \delta_i)\) for \((\hat{T}, L, \delta)\). Likelihood for a single observation \(i\) can be written using (1) as follows,

\[
L(\beta, \gamma_0|\mathbf{y}_i) = \prod_{s=1}^{l_i} \hat{\lambda}(s|\mathbf{x}_i)^{y_s}(1 - \hat{\lambda}(s|\mathbf{x}_i))^{(1-y_s)},
\]

where \(y_s\)'s are augmented pseudo-observations such that \(\{t_i, \delta_i = 1, l_i\} \Rightarrow \{y_{is} = 0; s = l_i, \ldots, t_i - 1; y_{il_i} = 1\}\) and \(\{t_i, \delta_i = 0, l_i\} \Rightarrow \{y_{is} = 0; s = l_i, \ldots, t_i - 1; y_{il_i} = 0\}\). Thus, (2) mimics a likelihood of binary logistic model with \(t_i - l_i + 1\) many observations.

### 2.1 Modeling hazard in presence of frailty

In fecundity studies, frailty is often introduced in modeling TTP to account for substantial unmeasured heterogeneity at subject-level.\(^7\) Representing frailty for \(i\)th subject by \(q\) random effects, say \(b_i\), where \(b_i \sim N(0, \mathbf{Q})\), discrete hazard, \(\hat{\lambda}(s|\mathbf{x}_i)\) from (2) can be given as follows,

\[
\hat{\lambda}(s|\mathbf{x}_i, \beta_0, \beta_i, \gamma_0) = h(\eta_s) = h(\gamma_s + \beta_0 + \mathbf{x}^T \beta + \mathbf{z}^T \mathbf{b}_i)
\]

\[
\eta_s = \gamma_s + \beta_0 + \mathbf{x}^T \beta + \mathbf{z}^T \mathbf{b}_i, s = 1, \ldots, t_i, i = 1, \ldots, n,
\]

where \(\mathbf{z}_s\) is the set of possibly time-varying covariates with random effects. In general, one can have \(\mathbf{x}_i\)'s vary over time. Let us define \(\mathbf{b}^T = (\mathbf{b}^T_1, \ldots, \mathbf{b}^T_n)\). Let \(\mathbf{A}_u\) be a binary vector of length \(t_{max}\) indicating position \(s\), and \(\mathbf{A}^T, \mathbf{X}^T, \mathbf{Z}^T\) be the corresponding matrices by stacking rows one over the other. Thus, \(\mathbf{b} \sim \mathbf{G}(\mathbf{b}) = N(0, \mathbf{Q}_b = I_n \otimes \mathbf{Q})\).

### 2.2 Variable selection with correlated covariates

Motivated by the fecundity study that measures environmental toxicants in blood which are often highly correlated as discussed in Section 1, we assume that covariates may be potentially highly correlated. Coupled with the fact that the number of covariates \(p\) is moderately large compared to sample size, correlated covariates pose an extra challenge in estimating the regression coefficients as standard methods would fail to produce stable and efficient estimates. This necessitates a variable selection strategy. We propose a penalized regression framework with elastic net penalty as, unlike Lasso, it gives equal weightings to all correlated variables and produces stable paths.\(^{10,11}\) To obtain robust estimate on baseline hazard, we apply a ridge penalty on \(\gamma\). Let \(l(\beta_0, \gamma, \mathbf{Q}|\mathbf{y})\) be log of marginal likelihood (or aka integrated) where \(\mathbf{y}' = [\mathbf{y}_1', \ldots, \mathbf{y}_n']\).
The penalized likelihood can be given using (3) as follows.

\[ l_{\text{pen}}(\beta_0, \beta, \gamma, Q|y) = l(\beta_0, \beta, \gamma, Q|y) - \nu \left( \alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 / 2 \right) - \nu_s \sum_{m=1}^{l_{\text{max}}} \gamma_m^2 / 2, \]

\[ \nu, \nu_s > 0, \quad 0 < \alpha < 1. \]  

(4)

where \((\nu \text{ and } \alpha)\) are penalty parameters for elastic net whereas \(\nu_s\) is the ridge penalty for the baseline parameters.

### 3 | ESTIMATION

Estimating parameters require special attention due to (a) presence of random effects as the integrated likelihood does not admit a closed analytical form, (b) non-differentiability of L1 penalty in elastic net. Let \(\theta\) denote the set of all parameters except frailty parameters, \(Q\).

#### 3.1 | Maximum likelihood (ML) estimation

Given a set of values for the tuning parameters, \(\nu, \alpha, \text{ and } \nu_s\), we propose to find ML estimates by coordinate wise gradient ascent approach for \((\theta, Q)\). We implement a (modified) gradient ascent approach\(^1^4\) for \(\theta\) fixing coordinates of \(Q\) whereas we block-update \(Q\) using EM-based method\(^1^7\) fixing coordinates of \(\theta\). We iterate between these updates until convergence. For ML estimation it is sufficient to work with the following approximated penalized likelihood with \(qn\) many new augmented parameters \(b\), that can be obtained from (4) by applying Laplace method on the integrated unpenalized likelihood part.\(^2^3\)

\[ l_{\text{pen}}^{\text{app}}(\beta_0, \beta, \gamma, b, Q|y) = \log f(y|\beta_0, \beta, \gamma) - 1/2 b^T Q_b^{-1} b - \nu \left( \alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 / 2 \right) - \nu_s \sum_{m=1}^{l_{\text{max}}} \gamma_m^2 / 2. \]  

(5)

Thus, we define \(\theta = \{\beta_0, \beta, \gamma, b\}\). We have summarized the optimization algorithm for \(\text{DisCnet}\), in detail for logit link in Algorithm 1 of Appendix B. The algorithm implements a path following mechanism on the grid of tuning parameters to find parameter estimates, and eventually, refits the model with the selected variables by optimizing (5) with elastic net penalty removed for post-selection inference.

#### 3.2 | Tuning and faster convergence

The parameters, \(\nu, \nu_s, \text{ and } \alpha\) require tuning together. We first estimate \(\nu_{\text{perm}}(\alpha, \nu_s)\) using permutation selection\(^1^9\) for a fixed pair of \((\alpha, \nu_s)\), where each such pair can be selected from a rectangular grid. The optimal combination can be chosen by minimizing the overall BIC(\(\nu_s, \alpha, \nu_{\text{perm}}(\nu_s, \alpha)\)) or prediction error of cross-validated samples. We recommend working with a reasonable set of grid points for \(\nu_s\), while we note that earlier studies indicate fixing \(\nu_s = 100\) to reduce computation time in similar real data applications.\(^1^8\) We have worked with a grid of \((15, 25, 50, 100)\) that was used both for simulation as well as for EDC mixture data applications. In Web Appendix G, supporting information, distribution of various tuned, and finally selected grid values are laid out across real data application, and simulation scenarios. As pointed out in Groll and Tutz,\(^1^8\) in general ridge penalty is assumed to produce robust estimates of baseline hazard parameters, which is particularly recommended in high dimensional settings where sample size is less or comparable to the parameters. While robust estimates may induce shrinkage, and thus may introduce bias in estimates, they are often treated as nuisance parameters. Further, in simulation studies they demonstrated little or no impact on variable selection accuracy. Once the variable selection is done, for post-selection inference the model with the selected variables is fitted using Fisher scoring (while still penalizing the baseline with ridge penalty; only elastic net penalty is dropped at this stage). In Algorithm 1, \(\theta\) update step (see 4. of Step 2) allows for a Fisher scoring step that can help speed up convergence, but is only applicable when optimal step size is within maximum allowed step-size, and the chain has gotten into a region where the gradient is continuous. Thus, Fisher scoring fails when the chain moves around the region of discontinuity. Under these circumstances,
it is recommended to let the gradient ascent run initially long enough until it lands into a region of continuous gradient, after which it can be replaced with iterated weighted least square Fisher’s scoring.

3.3 Adjustments for group selection with elastic net

We next present how our proposed methods can be adapted to group selection. We modify Meier et al’s\textsuperscript{24} approach to include a group specific $L_2$ norm penalty $||.||_2$ after we club coefficients $\beta$ into groups of variables. Let $\beta_g$ be a sub-vector of $\beta$ that corresponds to $g$th group. Let $m_g$ be the length of $\beta_g$. Then the penalized likelihood can be written as follows,

$$l_{pen}(\beta_0, \beta, \gamma, Q|y) = l(\beta_0, \beta, \gamma, Q|y) - \nu \sum_{g=1}^{G} (\sqrt{m_g}||\alpha \beta_g||_2 + (1 - \alpha)\beta_g^T \beta_g/2)$$

$$- \nu_s \sum_{m=1}^{l_{max}} \gamma_m^2/2, \quad \beta^T = (\beta_1^T, \ldots, \beta_G^T); \quad \nu, \nu_s > 0, \quad 0 < \alpha < 1.$$

Note when $g = 1$, $||.||_2$ becomes $L_1$ penalty. The estimation requires Algorithm 1 to modify $\beta$ update step. The changes are listed in Algorithm 2 as given in the Web Appendix B.

4 SIMULATION STUDY

We evaluate the effectiveness of our method, Discnet, in two different ways: Scenario I and Scenario II. In Scenario I, we compare its performance with glmmlasso (referred to as glmmlasso_discrete in Groll et al\textsuperscript{18}) in the context of frailty modeling for discrete survival analysis. Since our primary focus is to compare methods that deal with random effects models, there are not many options available to the best of our knowledge. Most of the other approaches, that allow variable selection, do not allow us to incorporate ridge penalty on baseline directly. Groll et al\textsuperscript{18} compared glmmlasso with approaches that allow random effects, such as glmex,\textsuperscript{25} gam4\textsuperscript{26} in generalized linear mixed model set-ups adapting them to stepwise variable selections. They also noted that in absence of random effects, glmmlasso performs at par with popular approaches, such as glmnet,\textsuperscript{27} penalized.\textsuperscript{14} Our proposed method implements elastic net penalty, which is critical for correlated variables, while additionally incorporating left truncated observations, as well as more efficient permutation-based tuning on $\nu$. In Scenario II, we evaluate effectiveness of group selection across scenarios, when multiple groups are present along with correlated variables, and with left truncated and right censored discrete survival time.

4.1 Scenario I: Impact of left truncation and random right censoring on performance

We simulate scenario similar to ones encountered in assessing environmental toxicants with survival data. We work with fixed set of $p = 150$ covariates in our simulation setup. We consider 12 cases based on three design parameters, namely, 1. $n$ = sample size with two levels (150 ($n = p$), 250 ($n > p$)), 2. $Cn$ = censoring proportion with three levels (low = 20%, medium = 35%, and high = 50%) and 3. $T_r$ = truncation indicator with two levels (0 = No truncation, 1 = 40% left truncation). We further assume $p^* = 5$ covariates are associated with discrete survival time $T$ such that $T$ takes values in $\{1, \ldots, T_{max}=10\}$. Thus, $T$ is simulated from the following frailty model:

$$\lambda_{i,t} = g(\gamma_t + \sum_{i=1}^{p} \beta_i * x_i + r_i); \quad r_i \sim N(0, \sigma^2 = 1), g = \text{logistic}(.)$$

in which the vector of non-zero regression coefficients, $\beta^* = (\beta_1, \ldots, \beta_5)' = (-4, -4, -4, 8, 8)'$, and $\beta_i = 0, p^* < i \leq p$. The time-varying baseline hazard parameters $\gamma$ is assigned: $(-9.00, -7.00, -4.97, -2.82, 0.34, 1.39, 2.35, 4.27, 6.18, 8.11)'$. $\gamma_t$ is chosen an increasing function of $t$ such that there is good representation of various instances of survival time, $T$, at all levels of censoring. In presence of left truncation ($T_r = 1$), left truncation variable, $L$, is simulated from
### Table 1: FP, FN, and MSE$_p$ (Med_SE) based on 1000 replications with correlated covariates ($R = 0.7l_1 + 0.3l_1 \odot 0.4l_2 + 0.6l_2$), for Discnet and glmlasso respectively $p = 150$

| Cases | Discnet | | glmlasso | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n  | Cn | Tr | FN | FP | Med_SE | FN | FP | Med_SE | |
| 150 | 0.21 | 0.34 | 0.00 (0.05) | 0.01 (0.11) | 100.55 (3.63) | 0.00 (0.06) | 0.31 (0.65) | 100.61 (3.70) | |
| 250 | 0.21 | 0.34 | 0.00 (0.00) | 0.01 (0.11) | 95.42 (2.72) | 0.00 (0.03) | 0.50 (0.95) | 95.45 (2.80) | |
| 150 | 0.36 | 0.33 | 0.02 (0.15) | 0.01 (0.09) | 98.31 (4.53) | 0.02 (0.17) | 0.29 (0.58) | 98.55 (4.69) | |
| 250 | 0.36 | 0.32 | 0.00 (0.00) | 0.01 (0.12) | 93.83 (3.34) | 0.00 (0.00) | 0.47 (0.88) | 93.92 (3.40) | |
| 150 | 0.50 | 0.30 | 0.09 (0.29) | 0.03 (0.19) | 93.24 (5.76) | 0.07 (0.28) | 0.30 (0.57) | 93.59 (6.08) | |
| 250 | 0.50 | 0.30 | 0.00 (0.06) | 0.02 (0.15) | 89.61 (4.41) | 0.00 (0.06) | 0.46 (0.81) | 89.73 (4.51) | |
| 150 | 0.22 | 0.00 | 0.00 (0.07) | 0.00 (0.06) | 101.66 (3.41) | 0.01 (0.08) | 0.26 (0.58) | 101.76 (3.47) | |
| 250 | 0.22 | 0.00 | 0.00 (0.00) | 0.00 (0.07) | 95.96 (2.55) | 0.00 (0.04) | 0.47 (0.91) | 95.96 (2.63) | |
| 150 | 0.38 | 0.16 | 0.03 (0.16) | 0.01 (0.09) | 98.87 (4.45) | 0.02 (0.15) | 0.30 (0.57) | 99.08 (4.58) | |
| 250 | 0.38 | 0.16 | 0.00 (0.00) | 0.01 (0.12) | 94.31 (3.29) | 0.00 (0.00) | 0.44 (0.85) | 94.32 (3.33) | |
| 150 | 0.53 | 0.28 | 0.09 (0.28) | 0.03 (0.16) | 92.91 (6.18) | 0.06 (0.28) | 0.28 (0.56) | 93.21 (6.44) | |
| 250 | 0.53 | 0.28 | 0.00 (0.07) | 0.02 (0.14) | 89.43 (4.42) | 0.01 (0.08) | 0.46 (0.81) | 89.65 (4.47) | |

Note: Discnet outperforms glmlasso consistently in terms of FP rates while FN rates remain comparable across three censoring levels; FN rates drop with the increase of $n$.

---

**Mult(0.6, 0.2, 0.2)** with labels 1, 2, 3 (with 1 indicating no truncation), motivated by real data as discussed in Section 5. Thus, for each choice of the triplet $(n, Cn, Tr)$, 1000 data replicates are generated from the model. Two approaches, namely Discnet and glmlasso are compared based on average performance over 1000 data-sets on three popular measurement metrics, namely average number of false negatives (FN), false positives (FP) and median squared error of estimator for coefficient vector $\hat{\beta}$ (Med_SE). One can obtain more commonly used quantities, such as false positive rate (FPR) and false negative rate (FNR) with FP and FN, divided by $p^*$ and $p - p^*$ respectively. While implementing glmlasso, optimal model is chosen based on BIC criterion. In case of implementation of Discnet, solution path is first constructed over $(\nu, \nu, \alpha)$, while $\alpha$ and $\nu$ are tuned over the grid $[1, 0.95, 0.8, 0.7, 0.6, 0.5] \times [15, 25, 50, 100]$ based on BIC. lasso parameter, $\nu$, is tuned following permutation approach. We have implemented this method under two settings, (1) with independent covariates, and (2) with correlated covariates. Under setting (1) we found that Discnet was as effective as that of glmlasso with improved performance in terms of FP rates and stability of results on some occasions. More details are given in Web Appendix C.

We elaborate the setting (2) here in which we generate correlated variables $x_i's$ in the following manner. We first generate independent $z_i's$ from $U(0, 1)$. We define $x_1 = z_1$, $x_2 = \theta_1 x_1 + z_2$, $x_3 = \theta_1 x_1 + z_3$. This implies $\text{cor}(x_1, x_2) = \text{cor}(x_1, x_3) = \frac{\theta_1}{\sqrt{1+\theta_1^2}} = \rho_1$ (therefore, $\theta_1 = \frac{\rho_1}{\sqrt{1-\rho_1^2}}$) and $\text{cor}(x_2, x_3) = \rho_1^2$. If we choose $\rho_1 = .9$, it corresponds to $\rho_1 \approx 2$. Continuing like this with $x_1$, one can define a block of variables (excluding $x_1$) with each pair having a fixed correlation $\rho_i^2$ (compound symmetry (CS)) leaving $x_1$. Thus, to define a CS with 2 blocks of size 3 each, and between-block independence, one can use $x_1 = \beta_1 z_{p+1} + z_t$, $i = 1, \ldots, 3$, and $x_1 = \beta_2 z_{p+2} + z_t$, $i = 4, 5, 6$. We make a note that although $z_i's$ are uniform $x'_i's$ behave more like triangular distribution ($U(0, 1) + U(0, 1) = \text{Triangular}(0, 2)$). In what follows, we have assigned 2-block correlation structure, each of size 3, with $\rho_1 = 0.7$ and $\rho_2 = 0.4$ in them, respectively, while the rest of $p - 6$ variables are generated from $U(0, 1)$. Out of these first 6 covariates, only first five are associated with the outcome variable as mentioned earlier. Performance metrics are reported in Table 1.

We again observe a similar pattern as in case 1 as our proposed method consistently remains as effective as the other, specially improving upon in terms of FP rates while remaining comparable with respect to other FN rates and MED_SE. From Table 1, it is evident that the proposed methods perform very well even in presence of truncation.
TABLE 2  FP, FN, non-group FN (NG_FN), MSE, percentage of GRP_1 captured (GRP1), and GRP_2 captured (GRP2) based on 1000 replicates with other correlated co-variates ($R = 0.7I_3 + 0.3I_5 \odot 0.4I_2 + 0.6I_2$) $p = 150$

| Type  | n  | Cn  | FN   | FP   | NG_FN | MSE  | GRP1 (%) | GRP2 (%) |
|-------|----|-----|------|------|-------|------|----------|----------|
| cont  | 150| 0.20| 0.90 | 0.05 | 0.81  | 231.08| 88.90    | 86.50    |
|       | 250| 0.20| 0.00 | 0.00 | 0.40  | 218.75| 100.00   | 100.00   |
|       | 150| 0.36| 3.11 | 0.24 | 0.68  | 239.50| 59.40    | 58.40    |
|       | 250| 0.36| 0.00 | 0.00 | 0.60  | 206.86| 100.00   | 100.00   |
|       | 150| 0.49| 4.62 | 0.43 | 0.56  | 321.81| 41.60    | 38.10    |
|       | 250| 0.49| 0.12 | 0.01 | 0.88  | 188.97| 99.10    | 97.50    |
| cat   | 150| 0.22| 0.09 | 0.09 | 0.67  | 230.19| 100.00   | 100.00   |
|       | 250| 0.22| 0.00 | 0.00 | 0.39  | 219.78| 100.00   | 100.00   |
|       | 150| 0.33| 0.17 | 0.16 | 0.76  | 225.46| 99.80    | 99.90    |
|       | 250| 0.33| 0.01 | 0.01 | 0.47  | 214.92| 100.00   | 100.00   |
|       | 150| 0.49| 0.52 | 0.37 | 0.86  | 206.45| 97.50    | 98.20    |
|       | 250| 0.49| 0.04 | 0.04 | 0.58  | 196.94| 100.00   | 100.00   |
| mixed | 150| 0.21| 0.27 | 0.02 | 0.75  | 232.03| 100.00   | 91.50    |
|       | 250| 0.21| 0.00 | 0.00 | 0.42  | 220.78| 100.00   | 100.00   |
|       | 150| 0.37| 1.17 | 0.12 | 0.61  | 230.12| 99.90    | 65.40    |
|       | 250| 0.37| 0.01 | 0.00 | 0.62  | 213.31| 100.00   | 99.80    |
|       | 150| 0.51| 2.42 | 0.37 | 0.51  | 262.87| 98.00    | 34.20    |
|       | 250| 0.51| 0.09 | 0.01 | 0.78  | 197.42| 100.00   | 97.20    |

4.2 Scenario II: Performance in group selection while other correlated covariates are present

In this scenario, our objective is to see how grouped elastic-net version performs across various scenarios. We borrow $\gamma$, $\beta^*$, Cn, and Tr from earlier set-up. Although, we kept the same correlation structure of 5 un-grouped variables, we add 7 variables, constituting two groups with first 4 and last 3 respectively of them, which are associated with the outcome. We consider three circumstances: (1) cont: both of the groups continuous in nature, (2) cat: both of them representing two categorical variables, and (3) mixed: the first group represents categorical one while the next one is continuous. Thus, if first one is a categorical variable, it has 5 levels with the 5th one being treated as reference class. As a result, we consider the following vectors of regression coefficients corresponding to two groups as follows:

$\beta_1^* = (7, -5, 7, -4)', \beta_2^* = (5, -8, 3)'$. To measure performance accuracy, we first define FP and FN as before based on all variables including the 7 new sub-group variables as variables in their own right. We further define a non-group false negative (NG_FN) if a non-group variable with non-zero coefficient is mis-classified as 0. We further compute group capture percentage for each of them based on 1000 replicates. All the metrics are reported in Table 2.

We first note that all groups are captured well when $n > p$. In general, average number of mis-classifications remains very small (<1) for non-group variables, and so is the case of FP. The same is also true for overall FN when censoring is at low level, 20%. We do observe, as in the earlier simulations, FN increases with the increase in censoring, Cn, while falls with the increase in $n$. We further note that when censoring is low, both types of groups are captured well. As censoring increases, group capturing percentage drops for the continuous group.

5 ANALYSIS OF LIFE STUDY: ENDOCRINE DISRUPTING CHEMICALS AND TIME-TO-PREGNANCY

In this section, we investigate the association of 46 (92 in total for both partners) endocrine disrupting chemical concentration (in gm/cm$^3$) from both partners on TTP, which were recorded in a prospective cohort study, namely The
Longitudinal Investigation of Fertility and the Environment (LIFE) Study, conducted between 2005 and 2009. The study targeted couples planning for pregnancy \((n = 501)\) who were followed for 12 months or until they got pregnant, or were lost to follow-up. In addition, couples were allowed to enroll within couple of cycles of going off contraception \((\text{max}(L) = 3, L - 1 \text{ cycles truncated})\). Thus, TTP is both right censored as well as left-truncated. The study measured various chemical exposures in blood, at the start of the study among which our focus lies on cotinine (indicator of smoking), lipid, and three classes of persistent organic pollutants (POP), including 9 organochlorine pesticides (OCPs, such as \(\beta\)-hexachlorocyclohexane (\(\beta\)-HCH)), polybrominated biphenyl (PBB-153), and 36 PCBs in addition to important risk factors, female age, delta age (difference between male and female partner’s age) and BMI \((\text{in kg/m}^2)\) for each partner of the couple. Our focus on these toxicants is motivated by previous epidemiological studies, that indicate single or individual class of toxicants’ exposure in pre-conception period may be adversely associated with fecundity, fetal growth and birth outcomes, such as birth weight, birth size and so forth.\(^{28,29}\) A detailed description of the summary statistics of the chemical toxicants including the value of limit of detection (LOD) and percentage below LOD can be found in Louis et al.\(^{30}\) However, studying it in the context of mixtures of chemical toxicants on fecundity has not been conducted. This was partly due to lack of available statistical methods to address this question. Performing a standard discrete survival analysis including all the toxicants poses few immediate challenges. As the number of variables, \(p = 100\), is quite large compared to sample size, and the variables are strongly correlated (see Figure 1), applying standard discrete frailty model may not produce efficient stable estimates, and thus necessitates a novel variables selection methodology, that addresses stable variable selection in discrete frailty modeling, in addition to right censoring and left truncation. In this context, elastic net penalization-based Discnet suits the purpose. This approach helps identify the “important drivers” of the mixtures of chemical toxicants.

Due to limited availability of biospecimens, chemical toxicants have missing values, which were imputed using standard methods, and 10 imputed copies of the data are analyzed separately to draw robust inferences. Since most of the chemicals register a low positive value, they are highly skewed to 0. For this reason, a log transform with offset 1\(^1\), \(f(x) = \log(1 + x)\), followed by standardization to have unit variance, is used to lessen skewness and stabilize the scale of the estimates. Thus a full candidate model, subjected to variable selection, can be given by,

\[
g(\lambda_{it}) = \gamma + \beta_0 + \beta_1 \text{Age}_t + \beta_2 \text{Delta}_\text{age} + \beta_3 \text{BMI}_t + \beta_4 \text{BMI}_m + \beta_5 \text{Cotinine}_t + \beta_6 \text{Cotinine}_m + \beta_7 \text{Lipid}_t + \beta_8 \text{Lipid}_m + \beta_9 \text{PCB028t} + \cdots + \beta_{100} \text{TNA}_t + b_i \quad \text{g(.)} = \logit, \quad \lambda_{it} \text{: hazard.}
\]

In addition, there are three other challenges that require special attention: accommodating—(a) lipid variable and its binding effect with some lipophilic chemicals in the model, (b) chemicals with high concentrations below LOD, and (c) potential non-linear effects of some chemicals. We discuss in detail next.

### 5.1 Accounting for lipophilic chemicals

Some chemicals, such as POP’s, tend to bind with lipid in blood, thereby higher presence of lipid would induce higher amount of the chemicals present in the blood due to binding. They are known as lipophilic chemicals. To give an example, free thyroid hormone is associated with higher levels of polybrominated diphenyl ether (PBDE) when PBDE was expressed on a lipid basis, but not when PBDEs were expressed on a wet-weight basis.\(^{31}\) In current literature on modeling lipids and lipophilic chemicals, often one of the two approaches is followed depending on whether a direct causal relationship can be attributed to lipid or not; (1) lipid as a covariate,\(^{32}\) (2) lipid as a concomitant variable.\(^{33}\)

While lipid as a covariate is straightforward to incorporate in the model, lipid as a concomitant requires new methods to be investigated when lipid is not assumed to be causally related. In the study of birth-defect, Li et al.\(^{33}\) suggested an adjustment to chemical-to-lipid ratio by replacing lipid with a Box-Cox transformation on it bringing in a power parameter, \(\kappa\) in the modeling of a binarized birth outcome. They pointed out further generalization to the adjustment may be required in a given problem. One drawback of their adjustment is that a known function of lipid is only allowed to act as a factor, flexibly inflating or deflating the chemical. In view of this, we propose slightly generalized Box-Cox transformation to the chemical-to-lipid ratio while modeling infertility, a binary outcome variable, as a first stage of a two-stage procedure. In the second stage, discnet is performed on the standardized transformed variables based on estimated \(\kappa\) with discrete survival outcome. Thus, if \(S_i\) is the lipid level of \(i\)th individual and \(X_i\) is the corresponding chemical concentration, we propose a two-stage procedure as follows.
FIGURE 1  Correlation globe showing the relationships of chemicals between females, males, and across other risk factors. Right-half represents females partners; left-half represents the male counterparts. Only Spearman’s rank correlations greater than 0.25 and smaller than −0.25 were shown as connections in the globe. Red line denotes positive correlation, and dark green (none observed) line denotes a negative one. Color intensity and line width are proportional to the size of the correlation. Within-class and between-class correlations are shown outside and inside of the track respectively.

\[
\begin{align*}
\text{Stage 1: Fit logit } (p_i) &= a_0 + a_1 x_i^*, x_i^* &= \text{standardized } g(x_i, s_i; \kappa), \text{ for each chemical using Bayesian logistic regression,} \\
g(x_i, s_i; \kappa) &= \text{Box-Cox}\left(\frac{\log(k_x + x_i)}{\log(1 + s_i)}, \kappa \right), \log(k_x + x_i) > 0, \\
\text{Box-Cox}(y; \kappa) &= \begin{cases} 
(0^\kappa - 1)/\kappa, & \kappa \neq 0 \\
\log(y), & \kappa = 0.
\end{cases}
\end{align*}
\]

Estimate \( \kappa \) by posterior median \( \hat{\kappa} \)

Stage 2: Fit Discnet with the transformed variables, standardized \( g(x_i, s_i; \hat{\kappa}) \) as defined in Stage 1 excluding lipid.

In above formulation \( p_i \) denotes probability of being infertile, a binary outcome variable. A couple is treated infertile if they do not get pregnant in next 12 months. The fitting methodology only tweaks Li et al’s procedure by defining a more generalized Box-Cox function. Every other aspect of MCMC computation implementing fast logistic regression remains the same. For the sake of completion, we outlined the MCMC steps in Web Appendix D in the supporting information.
5.2 Issue of limit of detection (LOD) and potential non-linear effects

Since most of chemicals are often found in blood in small proportion, their measurement suffers from what is known as limit (LOD) of detection issue. In other words, any registered value below a threshold, called LOD, may not be reliable. Though quite a many approaches were suggested in the literature for bias and variance correction for such noisy data, two popular approaches to treat them, that we have implemented here, are: (1) replace all the values below LOD by \( \frac{LOD}{\sqrt{2}} \) (referred to as “With LOD treatment” or WL), (2) keep all machine reported values as it is (referred to as WOL).

Apart from LOD, one of the drawbacks of our framework is the assumption of linear effects of covariates in log-odds ratio. In this context, some chemicals, long suspected of non-linearity, need to be dealt with differently. To address this, we again adopt two separate approaches, (1) without any non-linearity treatment, (2) binarizing potential non-linear covariate. To elaborate the second approach, we first fit penalized spline with each chemical in an unadjusted model, then identify chemicals based on how severe (significant) the estimated non-linearity is. We have used \texttt{coxph} function from \texttt{survival} package along with \texttt{pspline} function for non-linearity. We then converted them to binary variable based on a threshold, above which non-linearity is suspected. 22 such chemicals are found based on 10% significance level (See Web Appendix E for more details). Figure 2 represents two such chemicals, PCB087, DDT, for whom termplots are shown. The red vertical line dictates approximately the point above which non-linearity is perceived. These 22 variables are replaced by their binarized versions in the full model. In this regard, an alternative approach could be to use spline to model the mixture coefficients and implement a grouped elastic net to capture the relevant chemicals. Although this promises to be an appealing approach current methodology may not be best suited to extract weak non-linear relationships primarily due to the following challenges: significant increase in parameter complexity, and covariate matrix potentially becoming ill-conditioned. To get a sense of complexity, currently parameter complexity, which is \( T \) (baseline) + \( p \) (covariate) + \( n \) (sample size of unaugmented data), would increase to \( T + p \times K + n \), where \( K \) is the number of parameters required for a single spline fitting which further depends on the degree of the spline, and knot sizes. This would pose significant challenge in terms of implementing current algorithm specially in the context of low to moderate sample size problems, unless initial variable screening is performed. Further, knot size tuning is itself a penalized procedure that will only add to overloading of parameters. Because of these concerns, spline fitting for each covariate coefficient is currently considered out of scope. In the same vein, we also would like to add that to capture non-linearity, and to reduce the effect of various data transformations, we had considered further analysis representing chemicals by their tertiles, and applied a grouped elastic net grouping them at chemical level. This almost doubles the covariate size (excluding few chemicals for which tertiles did not make sense). We notice that tertiles corresponding to female cotinine got captured in addition to female age while other weak signals such as PCB101\textsubscript{m}, male cotinine level were missed. Thus, we believe capturing groups of weak signals requires larger sample size.

5.3 Variable selection and post-selection inference

So far, we have considered 8 scenario’s (non-linearity \( \times 2 \), lipid treatment \( \times 2 \), LOD \( \times 2 \)). Besides these, we further screen out chemicals which have high proportion of values below LOD and re-ran the 8 scenario’s for a smaller subset of 47 variables. The results from these analyses are contrasted with that of the full set of chemicals in the Web Appendix F. \texttt{Discnet} methodology is applied to the model with transformed variables, possibly preceded by a fitting procedure such as generalized Box-Cox model. \texttt{Discnet} is composed of two steps: \textit{Step 1}. Finalizing variable selection, \textit{Step 2}. Fit the model without penalizing chemicals, while penalizing baseline for robust estimation. In \textit{Step 1} since there are 10 imputed datasets, \texttt{Discnet} is performed on each of them, and set of chemicals selected in at least 6 of 10, are considered for refitting final model. \texttt{Discnet.R} can be implemented after specifying a set of grid points for tuning parameters. We have used same set of grid points, and tuning mechanism for \( \nu, \nu_s, \alpha \), as used in simulation settings (see Section 4).

In \textit{Step 2} we re-fitted the same model with the selected variables only, with no further penalization on them (except on baselines). We obtain the parameter estimates using iterated weighted least square, and estimated asymptotic variances for each of the imputed data. Eventually parameter estimates and their SDs from 10 imputations are summarized in a single table using \textit{Rubin’s} rule. Table 3 gives odds ratio and its 95% CI’s for scenarios prior to non-linearity treatment.

We note the female smoking behavior (cotinine levels) and female age are negatively associated with fecundity, and thus increase risk across all 4 modes of analysis, while male’s smoking behavior seems equally important as it seems to be associated in two sub-scenarios. Although quite a many PCB’s got selected across various modes of analysis, they may not be strongly associated with pregnancy risk. Among them, PCB028\textsubscript{f} seems to be positively associated with fecundity when
Figure 2: The attached termplots represent chemicals, that when kept in an unadjusted model, show significant non-linear effect ($P$-value < 0.1) in explaining time to pregnancy. Each of them is converted to a binary variable, with 1 indicating values crossing a threshold, above which non-linear effect is experienced.

Table 3: Post selection inference: Selected variables and estimated OR across various modes of analysis

| Linear effects assumed for chemicals | Lipid (covariate) | Lipid (concomitant) |
|-------------------------------------|-------------------|---------------------|
| Chem | WOL (0.12, 0.17) | WL (0.12, 0.17) | WOL (0.12, 0.18) | WL (0.12, 0.18) |
| (Intercept) | 0.15 | 0.15 | 0.14 | 0.14 |
| Cotinine$_f$ | 0.82 (0.7, 0.95) | 0.79 (0.68, 0.91) | 0.8 (0.69, 0.92) | 0.84 (0.72, 0.98) |
| AOE$_m$ | 0.9 (0.79, 1.03) | 0.9 (0.79, 1.03) | 0.89 (0.77, 1.01) | 0.89 (0.77, 1.01) |
| PCB049$_f$ | 1.05 (0.78, 1.42) | 1.13 (0.59, 2.18) | 1.14 (1.02, 1.28) | 1.15 (1.02, 1.29) |
| PCB028$_f$ | 1.07 (0.54, 2.1) | 1.07 (0.54, 2.1) | 1.07 (0.54, 2.1) | 1.07 (0.54, 2.1) |
| PCB028$_m$ | 1.04 (0.74, 1.45) | 1.04 (0.74, 1.45) | 1.04 (0.74, 1.45) | 1.04 (0.74, 1.45) |
| PCB101$_m$ | 1.13 (1, 1.28) | 1.13 (1, 1.28) | 1.13 (1, 1.28) | 1.13 (1, 1.28) |
| PCB044$_f$ | 1.11 (0.56, 2.23) | 1.11 (0.56, 2.23) | 1.11 (0.56, 2.23) | 1.11 (0.56, 2.23) |

Note: Each blank cell indicates non-selection of that chemical in Step 1.

treated as a function to ratio-to-lipid. PCB028$_m$ turns out to be positively associated when lipid is treated as a covariate, and no LOD treatment is done. Table 4, on the other hand lists similar figures when chemicals suspected of non-linearity are converted to binary variables.

We note that in this case, Discnet retains all variables as were selected in Table 3. We also observe odds ratio and their 95% CI levels remain almost the same. When compared with Table 4, Table 3 fails to pick any chemical from the list, suspected of non-linearity. On the other hand Table 4 selected two such chemicals, namely PCB194$_m$, DDT$_f$, one of which is well known harmful chemical, DDT$_f$, which came negatively associated with TTP across all 4 modes of analysis. Thus, Discnet is capable of selecting important chemicals while making robust inferences about them.
TABLE 4 Post selection inference: Chemicals, suspected of non-linear effect, are binarized based on a threshold, above which non-linearity is suspected.

| Suspected non-linear effects accommodated | Lipid (covariate) | Lipid (concomitant) |
|------------------------------------------|-------------------|---------------------|
| (Intercept)                              | WOL               | WL                  |
|                                           | 0.17 (0.14, 0.2)  | 0.16 (0.13, 0.19)  |
| Cotinine $f$                             | WOL               | WL                  |
|                                           | 0.82 (0.7, 0.95)  | 0.82 (0.7, 0.96)    |
| Cotinine $m$                             | WOL               | WL                  |
|                                           | 0.91 (0.79, 1.04) | 0.89 (0.77, 1.02)   |
| PCB194 $m$                               | WOL               | WL                  |
|                                           | 0.85 (0.64, 1.12) |                     |
| Age $f$                                  | WOL               | WL                  |
|                                           | 0.86 (0.75, 0.98) | 0.85 (0.75, 0.96)   |
| DDT $f$                                  | WOL               | WL                  |
|                                           | 0.75 (0.57, 0.98) | 0.75 (0.57, 0.98)   |
| PCB049 $f$                               | WOL               | WL                  |
|                                           | 1.06 (0.77, 1.47) | 1.17 (0.61, 2.25)   |
| PCB028 $f$                               | WOL               | WL                  |
|                                           | 1.02 (0.54, 1.9)  | 1.14 (1.01, 1.28)   |
| PCB028 $m$                               | WOL               | WL                  |
|                                           | 1.08 (0.69, 1.69) |                     |
| PCB044 $f$                               | WOL               | WL                  |
|                                           | 1.13 (0.58, 2.19) |                     |
| PCB101 $m$                               | WOL               | WL                  |
|                                           | 1.13 (1, 1.29)    |                     |

Note: Chemicals asterisked (*) represent the set of such cases, as selected by Discnet. Among them, only DDT* seem to be significantly negatively associated with the pregnancy across most of the cases. Chemicals marked †, on the other hand, represent cases having at least 50% of values above LOD.

6 | DISCUSSION

This article proposes a variable selection approach to identify the important drivers in the mixtures of endocrine disrupting chemicals on fecundity. In our analysis of the LIFE study, we identified female cotinine and female DDT as important drivers of the mixtures, with both of them negatively associated with fecundity. These findings are consistent with the previous literature that dealt with a subset of chemicals at a time. These findings were upheld over various modes of analysis. To the best of our knowledge, this is the first method that is applicable to assess mixture of chemical toxicants association to TTP, a discrete survival time subject to left truncation and right censoring.

We have proposed a novel variable selection, and fitting methodology, Discnet, in discrete hazard model with frailty via elastic net penalty for strongly correlated covariates, where survival time is both right censored and left-truncated. The elastic net-based method proposes a tuning mechanism that chooses $\lambda$ (lasso-part) penalty using permutation method, and while others using BIC. Importance of elastic net penalty to deal with high dimensional correlated covariates in discrete frailty modeling is highlighted through extensive simulations showcasing its effectiveness compared to glmlasso. Discnet improves the performance compared to its peers in presence of correlated covariates both in terms of stability and FP rates. It seems to work very well, in either $n > p$ scenario and/or low censoring cases, otherwise it performs at par with other methods. Additionally, we have extended our methodology to grouped selection to include categorical variables, and have shown the performance remains equally good. Finally, our methodology also differs from that of glmlasso in how the tuning parameters are chosen. We have used permutation selection method as it favors moderate sized models and a more balanced trade-off between power and false discovery rate. In the time to pregnancy study, we performed variable selection, and subsequent model fitting in multiple scenarios corresponding to multiple ways to deal with LOD issue, treatment of lipid in presence of lipophilic chemicals, as well as for the treatment of non-linear effects. In the process, we have suggested a new method based on generalized box-cox transformation that treats lipid as a concomitant variable.

Although in general, it is hard to establish asymptotic properties in our framework due to non-availability of closed analytical form of the likelihood, and involvement of multiple penalties, it is noted in earlier studies that L1-penalty based methods generally do variable selection well, but introduce bias in estimates via shrinkage. One can easily implement an adaptive version of our approach by first fitting a lasso/elastic net based estimates, and the subsequently
re-fitting the data with the elastic net penalty after scaling the covariates with the inverse of non-zero estimates obtained in the first stage (please note that although oracle property does not hold true for lasso/elastic net penalties due to shrunk estimators, variable selection still remains consistent). However, we chose a two-stage procedure where we do variable selection exercises followed by refitting the model with a fixed set of variables for two reasons: fast implementation and ease in handling missing data by combining multiple imputed data-sets (details of implementation explained previously). We intend to look at the adaptive version of our approach especially in the context of missing data in future.

There are several aspects which have room for improvements for future extensions. The hazard model can be easily extended to cloglog model. We are currently working to incorporate interactions between covariates with heredity constraints for discrete hazard model with frailty while performing variable selection. It would be interesting to see how various choices of loss functions impact the variable selection strategy. We plan to incorporate nonlinear effect of covariates in the variable selection strategy through a generalized additive model type frameworks.

As observed in the correlation analysis male and female EDC exposure are observed to be correlated with each other due to sharing of similar lifestyles and other characteristics. Although we have not explicitly assumed the grouping of partners based on the correlation structure, we have implicitly considered that via using elastic net penalty. One of the strengths of elastic net penalty is that, if both the correlated covariates are associated with the outcome, they would be selected together even though we do not implement a grouped elastic net. In fact, since elastic net penalty has an inherent grouping effect, grouped-elastic net is often less used compared to grouped-lasso. An alternative approach could be to explicitly define the groups of partners, and apply a grouped lasso/elastic net. However, the drawback of this method is that if there is a weak signal between one partner’s exposure to the survival outcome, the chemical exposures for both partners may not be selected at all even though other partner’s exposure may be strongly associated with the outcome. In contrast, an alternative approach of latent-class models where latent classes are based on male-female pair on infertility have been developed by Zhang et al. This approach was developed for a binary outcome, that is, infertility (yes/no) but not yet developed for the context of survival outcomes, and thus is potentially exciting alternative to build upon for future works.

In this article, we have focused on the frequentist approaches for variable selection, which among other desirable properties are easier to compute. Alternatively, computationally intensive Bayesian non-parametric methods such as Bayesian kernel machine regression with variable selection have been shown to perform better than L1-penalized methods, especially in the context of chemical mixtures with continuous outcomes. However, they need to be investigated more thoroughly in the context of survival outcomes subject to left truncation and right censoring.

ACKNOWLEDGEMENTS
This research was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (contracts N01-HD-3-3355, N01-HD-3-3356, and N01-HD-3-3358). The authors would like to acknowledge that this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT
Data available on request from the authors. Discnet algorithm is implemented in an R-function, Discnet.R with a host of auxiliary subroutines supporting it. They can be obtained from https://github.com/sahaAbh/Discnet.

ENDNOTES
1 In what follows some of the chemicals are referred to by their short hand notation, such as DDT for p,p'-dichlorodiphenyltrichloroethane (DDT), TNA for trans-nonachlor etc.
2 We note that an offset 1 leads to log(1 + x) ≈ x when x is small.
3 A concomitant is a non-confounding covariate which helps improve the precision of the estimate of interest when other variables are adjusted for it. As an example, height plays a role of concomitant variable when BMI is used as a surrogate for obesity instead of weight.

ORCID
Abhisek Saha https://orcid.org/0000-0003-2860-436X
Rajeshwari Sundaram https://orcid.org/0000-0002-6918-5002
REFERENCES
1. Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, et al. Endocrine-disrupting chemicals: an endocrine society scientific statement. Endocr Rev. 2009;30(4):293–342.
2. Kortenkamp A. Low dose mixture effects of endocrine disrupters: implications for risk assessment and epidemiology. Int J Androl. 2008;31(2):233–240.
3. Bergman A, Heindel JJ, Jobling S, Kidd KA, Zoeller RT. WHO (World Health Organization)/UNEP (United Nations Environment Programme). The State of the Science of Endocrine Disrupting Chemicals. Geneva: UNEP/WHO; 2012. Geneva, UNEP/WHO; 2013:296.
4. Mitro SD, Johnson T, Zota AR. Cumulative chemical exposures during pregnancy and early development. Current Environ Health Rep. 2015;2(4):367–378.
5. Woodruff TJ, Zota AR, Schwartz JM. Environmental chemicals in pregnant women in the United States: NHANES 2003–2004. Environ Health Perspect. 2011;119(6):878–885.
6. Gini C. Premières recherches sur la fécondabilité de la femme. Proceedings of the International Mathematical Congress; Vol. 2, August 11, 1924:889–892; North-Holland, Toronto.
7. Ecochard R. Heterogeneity in fecundability studies: issues and modelling. Stat Methods Med Res. 2006;15(2):141–160.
8. Lazarevic N, Barnett AG, Sly PD, Knibbs LD. Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: a review of existing approaches and new alternatives. Environ Health Perspect. 2019;127(2):026001.
9. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B (Methodol). 1996;58(1):267–288.
10. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B (Stat Methodol). 2005;67(2):301–320.
11. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. J Royal Stat Soc Ser B (Stat Methodol). 2007;69(4):659–677.
12. Rosset S. Following curved regularized optimization solution paths. Adv Neural Inf Proces Syst. 2004;17:1153–1160.
13. Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics. 2003;19(17):2246–2253.
14. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. Biom J. 2010;52(1):70–84.
15. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–395.
16. Fan J, Li R. Variable selection for Cox’s proportional hazards model and frailty model. Ann Stat. 2002;30(1):74–99.
17. Groll A, Tutz G. Variable selection for generalized linear mixed models by L1-penalized estimation. J Royal Stat Soc Ser B (Methodol). 2014;76(2):305–338.
18. Groll A, Tutz G. Variable selection in discrete survival models including heterogeneity. Lifetime Data Anal. 2017;23(2):305–338.
19. Sabourin JA, Valdar W, Nobel AB. A permutation approach for selecting the penalty parameter in penalized model selection. Biometrics. 2015;71(4):1185–1194.
20. Buck Louis GM, Schisterman EF, Sweeney AM, et al. Designing prospective cohort studies for assessing reproductive and developmental toxicity during sensitive windows of human reproduction and development—the LIFE Study. Paediatr Perinat Epidemiol. 2011;25(5):413–424.
21. Cox DR. Regression models and life-tables. J Royal Stat Soc Ser B (Methodol). 1972;34(2):187–202.
22. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons; 2011.
23. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88(421):9-25. doi:10.1080/01621459.1993.10594284
24. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. J Royal Stat Soc Ser B (Stat Methodol). 2008;70(1):53–71.
25. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823, 2014.
26. Wood S, Scheipl F, Wood Maintain R. Package ‘gamm4’. 2008;31(2):233–240.
27. Abdelouahab N, Langlois M-F, Lavoie L, Corbin F, Pasquier J-C, Takser L. Maternal and cord-blood thyroid hormone levels and exposure to polybrominated diphenyl ethers and polychlorinated biphenyls during early pregnancy. Am J Epidemiol. 2013;178(5):701–713.
28. Schisterman EF, Whitcomb BW, Buck Louis GM, Louis TA. Lipid adjustment in the analysis of environmental contaminants and human health risks. Environ Health Perspect. 2005;113(7):853–857.
29. Li D, Longnecker MP, Dunson DB. Lipid adjustment for chemical exposures: accounting for concomitant variables. Epidemiology. 2013;24(6):921–928.
30. Robledo CA, Yeung E, Mendola P, et al. Preconception maternal and paternal exposure to persistent organic pollutants and birth size: the LIFE study. Environ Health Perspect. 2015;123(1):88–94.
31. Liu Z, Sun Z, Liu L, Li G. Broken adaptive ridge regression and its asymptotic properties. J Multivar Anal. 2018;168:334–351.
32. Zhang B, Chen Z, Albert PS. Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. Biostatistics. 2012;13(1):74–88.
38. Lazarevic N, Knibbs L, Sly P, Barnett A. Performance of variable selection methods for estimating the non-linear health effects of correlated chemical mixtures: a simulation study. *Environ Epidemiol*. 2019;3:224-225.
39. Bobb JF, Valeri L, Claus HB, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*. 2015;16(3):493-508.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Saha A, Sundaram R. Variable selection for discrete survival model with frailty in presence of left truncation and right censoring: Studying association of environmental toxicants on time-to-pregnancy. *Statistics in Medicine*. 2023;42(2):193-208. doi: 10.1002/sim.9609

---

**APPENDIX A. TREATMENT OF A CONCOMITANT VARIABLE IN A SURVIVAL MODEL**

A concomitant variable is defined to be a variable that may not directly influence an outcome but may influence the relationship of a covariate with the outcome variable. For example, Li et al.\(^{33}\) treated lipid as a concomitant variable which may influence the relationship between lipophilic chemicals and a health outcome of interest. In the context of survival model, the link function on discrete hazard is modeled as,

\[
\eta_{i,t} = \gamma_0 + \sum_{j=1}^{p} \beta_j \ast f_j(x_j, s) + b_t,
\]

where \(s\) is treated as a concomitant variable which impacts the relationship of covariate \(x_j\)'s with the TTP—captured by the function \(f_j\). Li et al used \(f_j(x_j, s) = x_j / g(s; \kappa_j)\), \(g(s; \kappa_j) = \text{Box-Cox}(s, \kappa_j)\) in their study with lipids, and estimated \(\kappa_j\)'s using a Bayesian logistic model (note that a binary health outcome was modeled in their study). In Section 5, while treating lipid as a concomitant we worked with slightly more generalized Box-Cox function (as defined in Section 5) for \(f_j(x_j, s)\) and implemented Bayesian logistic model treating infertility as health outcome to obtain estimates for \(\kappa_j\)'s to be used in second stage analysis of variable selection and association study.
APPENDIX B. MODIFIED GRADIENT ASCENT WITH EM TYPE UPDATES

Algorithm 1. Coordinate-wise gradient (modified) ascent

For a fixed set of values of the tuning parameters, $v$, $\alpha$, and $v_i$,

1. **Definitions:**
   1. $\theta$: Set of all parameters except $Q$
   2. $H^T := (X^T A^T Z^T)$
   3. $\lambda_i(\theta)^T = (\lambda(s = 1|x_1; \theta), \ldots, \lambda(s = t_1|x_1; \theta), \ldots, \lambda(s = t_q|x_q; \theta))^T$.

2. **Step 1:** Initialize parameters of interest, $(\theta = \theta^{(0)} = (\beta_0^{(0)}, \beta^{(0)}, \gamma^{(0)}, b^{(0)}), Q = Q^{(0)})$
   Iterate over $l = 1, 2, \ldots$, until convergence

   - **Step 2:** Block-update $\theta = \theta^{(l)}|\hat{Q}^{(l-1)}, Q^{(l-1)}$ as follows,
     1. Compute modified gradient of penalized likelihood, $S_{\text{pen}}(\theta)$ at $\theta^{(l-1)}$.
        With $S(\theta) = \partial(\theta)/\partial \theta = H[y - \lambda(\theta)]$,
        $S_{\text{pen}}^{(0)}(\theta^{(l-1)}) = S_0(\theta^{(l-1)})$,
        $S_{i}^{\text{pen}}(\theta^{(l-1)}) = \begin{cases} S_i(\theta^{(l-1)}) - v(1 - \alpha)\beta_i^{(l-1)} - \alpha \times \text{sign}(\beta_i^{(l-1)}), & \text{if } \beta_i^{(l-1)} \neq 0 \\
        S_i(\theta^{(l-1)}) - \alpha \times \text{sign}(S_i(\theta^{(l-1)})), & \text{if } \beta_i^{(l-1)} = 0 \text{ and } |S_i(\theta^{(l-1)})| > \alpha \\
        0, & \text{Otherwise}
        \end{cases}$
        $S_{i}^{\text{pen}}(\theta^{(l-1)}) = S_i(\theta^{(l-1)}) - \beta_i^{(l-1)} - \beta_i^{(l-1)} - Q_{i}^{(l-1)}b$, $i = p + 1 + f; j = 1, \ldots, t_{\text{max}},$
        $\text{sign}(\alpha) = I(x > 0) - I(x < 0)$

   2. Compute the directional second derivative at any $\theta$ and at any direction $v$, $I_{\text{pen}}(\theta, v) = -v^T F_{\text{pen}}(\theta) v$, where $F_{\text{pen}}(\theta) = -E(\nabla^2 \lambda_{\text{pen}}(\theta)) = \text{Fisher’s matrix}$. Note,
        $F_{\text{pen}}(\theta) = HH^T + K, K = \text{Diag}(v(1 - \alpha)I, v_3, Q_{b})^{-1}$,
        $W = \text{Diag}(., h(\eta_j)(1 - h(\eta_j)), .)$ wherever $S_{\text{pen}}$ is differentiable;

   3. **Step 3:** Step size for gradient ascent step: compute $t_{\text{edge}}$ and $t_{\text{min}}$ as in Goeman14
      $t_{\text{edge}} = \min \left\{ -\frac{S_{i}^{\text{pen}}(\theta^{(l-1)})}{S_i(\theta^{(l-1)})} : \text{sign}(\beta_i^{(l-1)}) = -\text{sign}(S_{i}^{\text{pen}}(\theta^{(l-1)})) \neq 0 \right\}$,
      $t_{\text{opt}} = -\frac{|S_{i}^{\text{pen}}(\theta^{(l-1)})|}{|S_i(\theta^{(l-1)})|}$, $|.|_2$ being $L_2$ norm

   4. **Step 4:** Update $\theta \rightarrow \hat{\theta}^{(l)}$ by choosing optimal step without changing signs in any coordinate
      $\theta^{(l)} = \begin{cases} \hat{\theta}^{(l-1)} + \epsilon t_{\text{edge}} S_{i}^{\text{pen}}(\theta^{(l-1)}) & \text{if } t_{\text{opt}} \geq t_{\text{edge}} \\
      \hat{\theta}^{(l)} & \text{if } t_{\text{opt}} < t_{\text{edge}} \text{, and sign}(\theta^{(l)}) = \text{sign}(\theta^{(l-1)}), \\
      \hat{\theta}^{(l-1)} + \epsilon t_{\text{opt}} S_{i}^{\text{pen}}(\theta^{(l-1)}) & \text{otherwise}
      \end{cases}$
      $\text{sign}(\hat{\theta}^{(l-1)}) = \lim_{\epsilon \to 0} \text{sign}(\hat{\theta}^{(l-1)} + \epsilon S_{i}^{\text{pen}}(\theta^{(l-1)})), \hat{\theta}^{(l-1)} = \hat{\theta}^{(l-1)} + [F_{\text{pen}}(\theta) - S_{\text{pen}}(\theta^{(l-1)})].$

5. **Step 3:** Block-update $Q \rightarrow \hat{Q}^{(l)}|\hat{\theta}^{(l)}$, $\hat{Q}^{(l-1)}$ as follows, using EM-based update strategy, 17
      $\hat{Q}^{(l)} = \sum_{n=1}^{n} \left( V_{ii} + b_{i}^{(l)}b_{i}^{(l)} \right)^{-1} n, V_{ii} = F_{ii}^{-1} + F_{ii}^{-1} F_{i0} \left( F_{00} - \sum_{n=1}^{n} F_{0i} F_{i0} \right) ^{-1} F_{0i} F_{i0}^{-1}$
      $F_{ii}$ is a submatrix of $F_{\text{pen}}$ that corresponds to rows that represent $x$ and columns that represent $y$. In this case $\hat{\theta}^{T} = (\gamma^{T}, \beta_{0}, \beta_{\text{active}})$ (See Web Appendix A for more details).