Synthetic Data Set #1

1. Information provided on the website:

Chemical Mixture Simulated Data

These synthetic data can be considered as the results of a prospective cohort epidemiologic study. The outcome cannot cause the exposures (as might occur in a cross-sectional study). Correlations between exposure variables can be thought of as caused by common sources or modes of exposure. The nuisance variable Z can be assumed to be a potential confounder and not a collider.

Structure of data file:

Name:	DataSet1.xls
Format:	Excel file; the first row is a header, each row represents a subject
Number of records:	500
Data per subject:	$Y, X_1, X_2, X_3, X_4, X_5, X_6, X_7, Z$
	Y = outcome data, 1 continuous variable
	X_1 - X_7= exposure data, 7 continuous variables
	Z: potential confounder, binary

Additional information:

For purposes of Data Set #1, there is no loss to follow up, missing or censored data, mis-measurement of the variables (Y, X_i, Z), or many of the other potential biases. One may also assume that the seven exposure variables X_1 –X_7 and Z are not intermediate variables and not colliders. There are no other confounders or effect measure modifiers. Random noise has been added to the outcome variable.

2. Creation of dataset #1: Detailed Information

a. Joint distribution of exposures

The seven exposure variables were constructed to be approximately log normal. They are not quite log-normal as the data were truncated at the high end (X_i>5) to eliminate extreme points.

The means and standard deviations on the natural log-scale used to generate the data are

	mean	SD
$\log(X_1)$	0	1
$\log(X_2)$	0	0.5
$\log(X_3)$	0	1
$\log(X_4)$	0	0.7
$\log(X_5)$	0	1
$\log(X_6)$	0	0.8
$\log(X_7)$	0	1

The correlation matrix used to generate the data is as follows:

	$\log(X_1)$	$\log(X_2)$	$\log(X_3)$	$\log(X_4)$	$\log(X_5)$	$\log(X_6)$	$\log(X_7)$
$\log(X_1)$	1	0.9	0.9	0.4	0.1	0.1	0.3
$\log(X_2)$			0.9	0.4	0.1	0.1	0.3
$\log(X_3)$				0.4	0.1	0.1	0.3
$\log(X_4)$					-0.1	-0.1	0
$\log(X_5)$						0.7	0.2
$\log(X_6)$							0.2

X_1, X_2, X_3 are all strongly correlated (ρ=0.9); X_5 and X_6 are a bit less correlated (ρ=0.7). For example, such correlations might be caused by common sources among the clusters.

b. Covariate Z

As stipulated, Z is to be treated as a confounder and not a collider. It was constructed to be associated with Y and X_1 (and thus also with X_2 and X_3). As shown below, no effect measure modification was built into Z.

c. Dose-response function for the outcome Y

The outcome was generated as follows:

$$Y = f[X_1, X_2, X_4, X_5, X_7] + \gamma Z + \varepsilon$$

where $f[\ldots]$ is the function relating the outcome to the exposures, Z is the confounder (with constant γ) and ε is normally distributed noise:

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$f[\ldots]$ is a biologically-based dose response function based on endocrine disruption:

$$f[X_1, X_2, X_4, X_5, X_7] = \alpha_0 + \frac{\alpha_1\left(\frac{T}{K_T} + \frac{X_1}{K_1} + \frac{X_2}{K_2}\right)}{\left(1 + \frac{T}{K_T} + \frac{X_1}{K_1} + \frac{X_2}{K_2} + \frac{X_4}{K_4} + \frac{X_5}{K_5}\right)} R_0(X_7)$$

$$R_0(X_7) = R_{00} + \frac{\lambda X_7}{K_7 + X_7}$$

X_3 and X_6 had no influence on the outcome. These two variables were part of the two exposure clusters (X_1, X_2, X_3) and (X_5, X_6). This is biologically plausible as different congeners of a group of compounds (e.g., PCBs) may be highly correlated, but have different biological effects, both qualitatively and quantitatively. A challenge for analysts is to determine this, given the high correlation for the (X_1, X_2, X_3) cluster in particular.

Biological rationale: Consider the X_i as biomarkers of exposure and the effect Y as an endocrine-related outcome. *T* is an endogenous hormone with a constant concentration that binds to receptor. X_1 and X_2 are agonists, i.e., they also bind and activate the receptor. In contrast, X_4, X_5 are competitive antagonists. X_1, X_2, X_4, X_5 can be thought of as endocrine disruptors acting at the receptor level. This model, based on classical pharmacology, is non-linear and has exposures that act in opposite directions: X_1 and X_2 will be positively associated with the outcome, while X_4, X_5 will be negatively associated with the outcome. The parameters K_1, K_2, K_4, K_5 can be thought of as EC_{50} for agonists

or related to IC_{50} for antagonists, when compounds are examined individually (K_T is the EC_{50} for the endogenous ligand, but since T is constant, only the ratio T/K_T matters here).

Pharmacologic models often treat the concentration of receptors (R_0) as fixed. However, there are cases where compounds can modulate the receptor without being a ligand: X_7 is such a compound. The second equation in the model shows that it changes the receptor level from the baseline R_{00} (when X_7=0) via a simple Hill function with a Hill coefficient (exponent) of 1. The effect of X_7 on R_0 is half maximum when X_7=K_7 and reaches a maximum value of λ when X_7>>K_7.

d. Parameter values:

The following parameter values were used:

T/K_t	0.5
K_1	1.5
K_2	3
K_4	4.5
K_5	1
K_7	0.5
R_{00}	1
λ	2

Among the two agonists, X_1 is twice as potent as X_2. Among the two competitive antagonists, X_5 is 4.5 times as potent as X_4. Figure 1 shows the dose-response curves for the 5 exposure variables, with other exposures (and Z) set to zero. As shown, X_1, X_2, X_7 increase Y, while X_4, X_5 decrease Y. Although none of the curves are linear, X_7 is the most non-linear while X_4 looks the least.

Other parameters:

α_0	2	adjusts minimum of model
α_1	20	adjusts minimum of model
γ	10	strong confounder
σ_ε	2.32	amount of random noise

3. Sample answers to the posted questions:

1. Which exposures contribute to the outcome? Are there any that do not? (Qualitative)

Contribute: X_1, X_2, X_7 (positively); X_4, X_5 (negatively)

Don't contribute: X_3 and X_6

2. Which exposures contribute to the outcome and by how much? (Quantitative)

The full answer to this question is provided by specifying the dose-response function (or an approximation) and its parameters. A partial answer might be that 1) X_1 and X_2 are positively associated with the outcome with X_1 twice as potent as X_2; 2) X_4 and X_5 are negatively associated with the outcome with X_5 4.5 times as potent as X_4.

3. Is there evidence of "interaction" or not? Be explicit with your definition of interaction (toxicologists, epidemiologists and biostatisticians tend to think about this quite differently).

In terms of toxicology, there are the following kinds of interaction (relative to concentration addition):

X_1 and X_2	TEF (toxic equivalent factor), a special case of concentration addition (both increase Y)	
X_1 and X_4	competitive antagonism	(similarly for X_2 and X_4)
X_1 and X_5	competitive antagonism	(similarly for X_2 and X_4)
X_1 and X_7	supra-additive ("synergy")	(similarly for X_2 and X_7)
X_4 and X_5	TEF, a type of concentration addition (both decrease y)	
X_4 and X_7	antagonism (unusual kind)	(similarly for X_5 and X_7)

Statistical interaction will depend on how a model is constructed. Formal epidemiologic analysis is judged on the additive scale.

4. What is the effect of joint exposure to the mixture? (Qualitative)
One answer might be: X_1, X_2, X_7 are positively associated with the outcome; X_4, X_5 are negatively associated with the outcome

5. What is the joint dose-response function? For example, if you can describe Y as a function of the exposures, what is your estimate of the function $Y=f(X_1,\ldots,X_p)$? (Quantitative)

$$f[X_1,X_2,X_4,X_5,X_7]=\alpha_0+\frac{\alpha_1\left(\frac{T}{K_T}+\frac{X_1}{K_1}+\frac{X_2}{K_2}\right)}{\left(1+\frac{T}{K_T}+\frac{X_1}{K_1}+\frac{X_2}{K_2}+\frac{X_4}{K_4}+\frac{X_5}{K_5}\right)}R_0\left(X_7\right)$$

$$R_0\left(X_7\right)=R_{00}+\frac{\lambda X_7}{K_7+X_7}$$

α_0	2
α_1	20
T/K_t	0.5
K_1	1.5
K_2	3
K_4	4.5
K_5	1
R_{00}	1
λ	2
K_7	0.5

6. Provide metrics for your answer. For example, consider adjusted r square or root mean square error, etc.
This will depend on the model used. The error built into the model is normal with σ_ε =2.32.

7. Analysts may also provide a description of the joint distribution of the exposure data.

The exposure data are approximately log-normal with a correlation structure as described earlier.

	$\log(X_1)$	$\log(X_2)$	$\log(X_3)$	$\log(X_4)$	$\log(X_5)$	$\log(X_6)$	$\log(X_7)$
$\log(X_1)$	1	0.9	0.9	0.4	0.1	0.1	0.3
$\log(X_2)$			0.9	0.4	0.1	0.1	0.3
$\log(X_3)$				0.4	0.1	0.1	0.3
$\log(X_4)$					-0.1	-0.1	0
$\log(X_5)$						0.7	0.2
$\log(X_6)$							0.2

Fig 1. Dose response curves for the exposure variables, setting the other variables to zero.