

ExposureAssessment

A program for Exposure Assessment from continuous data in the absence of a gold standard test

(Version 1.4.1, April 2012)

1. Introduction

ExposureAssessment is a software package for estimating exposure probabilities in a series of subjects where one or more continuous test measurements relating to the exposure of interest are available for each subject. It is an implementation of the Bayesian latent class hierarchical model presented in section 3.2 of

Weichenthal S, Joseph L, Bélisle P, Dufresne A.

Bayesian estimation of the probability of asbestos exposure from lung fiber counts.

Biometrics 2010;66(2):603-612.

This paper is available from

<http://www.medicine.mcgill.ca/epidemiology/Joseph/publications/Methodological/weichenthal2009.pdf>

Typically, one or more data points are taken from a sample of subjects, and are subsequently used to estimate the probability of exposure to a substance of interest (e.g., a carcinogen such as asbestos fibres in the lung). This can be accomplished by comparing the values from each subject to the distribution of test results from unexposed and/or exposed populations.

ExposureAssessment is useful in providing individual level probabilities of exposure based on available data or to study the test properties (mean and variance of exposed and unexposed populations) of the various test measures at hand.

Since **ExposureAssessment** is based on Bayesian latent class models, it can analyze data and provide probabilities of exposure even when no perfect gold standard test measure is available. The software can also accommodate test measures with mixed discrete/continuous distributions.

Depending on the exposure under study, contaminants can be present in some samples but still below a given detection limit; if the contaminant measurement is normally or log-normally distributed (when detectable), then its overall distribution can be modeled as a mixed discrete/continuous distribution, that is, by a normal or log-normal distribution to which a point probability mass is attributed to below-detection values. Variables without minimum detection limits can also be analyzed by this software provided their distribution is normal or log-normal in both exposed and unexposed populations. This distribution, implemented within a hierarchical Bayesian latent class model, forms the basis for the probabilities calculated in **ExposureAssessment**.

2. Hierarchical model

Full details of the model used by **ExposureAssessment** are given in the reference from section 1, which should be read before the software is used. Briefly, the model fit by **ExposureAssessment** can be described as follows.

A series of V exposure variables are measured on a set of N subjects, where the number of measurements per variable can vary from subject to subject. Let X_{ijk} be the k th measurement of variable j in subject i and let $\{I_i\}_{i=1,\dots,N}$ be the true status for each subject, where

$$I_i = \begin{cases} 1 & \text{if subject } i \text{ was exposed} \\ 0 & \text{if subject } i \text{ was not exposed} \end{cases}, \quad i=1,2,\dots,N.$$

The values of each of the V exposure variables are modeled as a mixture of a normal density and a probability of being at or below the detection limit, so that

$$X_{ijk} \sim \begin{cases} N(\mu_{ij}^{(I_i)}, \sigma_{wj}^{2(I_i)}) & \text{with probability } 1 - p_j^{(I_i)} \\ \leq \varepsilon_j & \text{with probability } p_j^{(I_i)} \end{cases}$$

$i=1,2,\dots,N, \quad j=1,2,\dots,V, \quad k=1,2,\dots,n_{ij}$

where $\mu_{ij}^{(1)}$ and $\mu_{ij}^{(0)}$ are the individual means for variable j in exposed and unexposed populations, respectively, and $\sigma_{wj}^{2(1)}$ and $\sigma_{wj}^{2(0)}$ are the within-subjects variances for variable j in exposed and unexposed populations, respectively

The values $\{\varepsilon_j\}_{j=1,2,\dots,V}$ are the detection limits and the values $\{p_j^{(g)}\}_{j=1,2,\dots,V}$ are the at or below-detection probabilities in the exposed ($g=1$) and unexposed ($g=0$) populations.

The individual means are modeled through the hierarchical model

$$\mu_{ij}^{(g)} \sim N(\mu_j^{(g)}, \sigma_{Bj}^{2(g)}), \quad i=1,2,\dots,N \quad j=1,2,\dots,V, \quad g=0,1$$

where $\sigma_{Bj}^{2(g)}$, $g=0,1$, are the between-subjects variances for variable j and the parameters $\mu_j^{(g)}$ are the overall means for variable j in both exposed ($g=1$) and unexposed ($g=0$) populations and are modeled as

$$\mu_j^{(g)} \sim N(\mu_j^{*(g)}, \sigma_j^{2*(g)}), \quad j=1,2,\dots,V, \quad g=0,1.$$

The individual means for each of the V variables are subject to the constraints

$$\text{Sign}(\mu_{ij}^{(1)} - \mu_{ij}^{(0)}) = \delta_i, \quad i=1,2,\dots,N, \quad j=1,2,\dots,V$$

where

$$\delta_j = \begin{cases} 1 & \text{if variable } j \text{ is expected to take larger values in the Exposed population} \\ -1 & \text{if variable } j \text{ is expected to take larger values in the Unexposed population.} \end{cases}$$

The within- and between-subject variances for each of the V variables are given uniform prior distributions

$$\begin{aligned} \sigma_{Wj}^{(g)} &\sim U(\sigma_{WLj}^{(g)}, \sigma_{WUj}^{(g)}), & j=1,2,\dots,V, & \quad g=0,1 \\ \sigma_{Bj}^{(g)} &\sim U(\sigma_{BLj}^{(g)}, \sigma_{BUj}^{(g)}), & j=1,2,\dots,V, & \quad g=0,1 \end{aligned}$$

while the below-detection probabilities are given Beta prior distributions:

$$p_j^{(g)} \sim \text{Beta}(\alpha_{\epsilon_j}^{(g)}, \beta_{\epsilon_j}^{(g)}), \quad j=1,2,\dots,V, \quad g=0,1.$$

Finally, in subjects where Exposure status is unknown, the latent true status are Bernoulli with probability of being positive equal to the prevalence of exposure π in the population under study:

$$\begin{aligned} I_i &\sim \text{Bernoulli}(\pi), & i=1,2,\dots,N \\ \pi &\sim \text{Beta}(\alpha, \beta). \end{aligned}$$

Table 3 of Section 5 summarizes the above notation.

3. Data preparation

The data to be analyzed by **ExposureAssessment** must be available in comma-separated values (.csv) files.

All data available for unclassified subjects (that is, subjects whose exact exposure is not known with certainty) must be saved in a unique file, while (if available) data for known Exposed and Unexposed subjects will be saved to two separate files. Thus, **ExposureAssessment** will read in data from one, two or three comma-separated values input files, depending on the type of data available for analysis.

Each column in the first row of each input data file must consist of the corresponding column variable name. In other words, the first row in the input file must be a header row.

The **ExposureAssessment** graphical user interface allows the user to pick a subset of the variables in the input files for analysis. Thus, not every variable present in input data files needs to be included in the analysis.

Each data file may optionally contain a subject identity number, which may make data entry of the multiple measurement entries for each subject easier, and make the reference to each subject easier in the output file (where subject ID number will be displayed).

Multiple values for a same variable can be entered in as many columns as necessary, as long as the columns are labelled with the same variable name (note that **ExposureAssessment** is case insensitive with regards to variable names).

In the example below, the variable *short fibers* was measured four times for subject A-100 (two columns on each of this subject’s two rows) while it was measured three times for both subjects A-101 (two measurements entered in row 4, one in row 5) and A-102 (all three measurements on same row 6).

	A	B	C	D	E
1	IDNumber	short fibers	short fibers	short fibers	long fibers
2	A-100	520	530		80
3	A-100	600	505		
4	A-101	105	155		90
5	A-101	230			
6	A-102	450	800	610	105
7	A-103				780
8					
9					

Note that blank cells indicate missing values. Hence subject A-103 did not have any *short fibers* measurements in example above.

In files where no Subject ID variable is defined, it will be assumed that each line of data indicates a unique subject. In output sections where individual data or individual exposure probabilities are printed, the Subject ID variable (if defined) will be used as a label to identify each subject.

If no Subject ID variable is ever defined (so it does not appear in the unclassified, unexposed or exposed subject data files), subjects will by default be labelled as “Unclassified”, “Unexposed” or “Exposed” (respectively) followed by a dash and the corresponding input data file row number. If Subject ID variables are given in some files but not in others, the above default labels

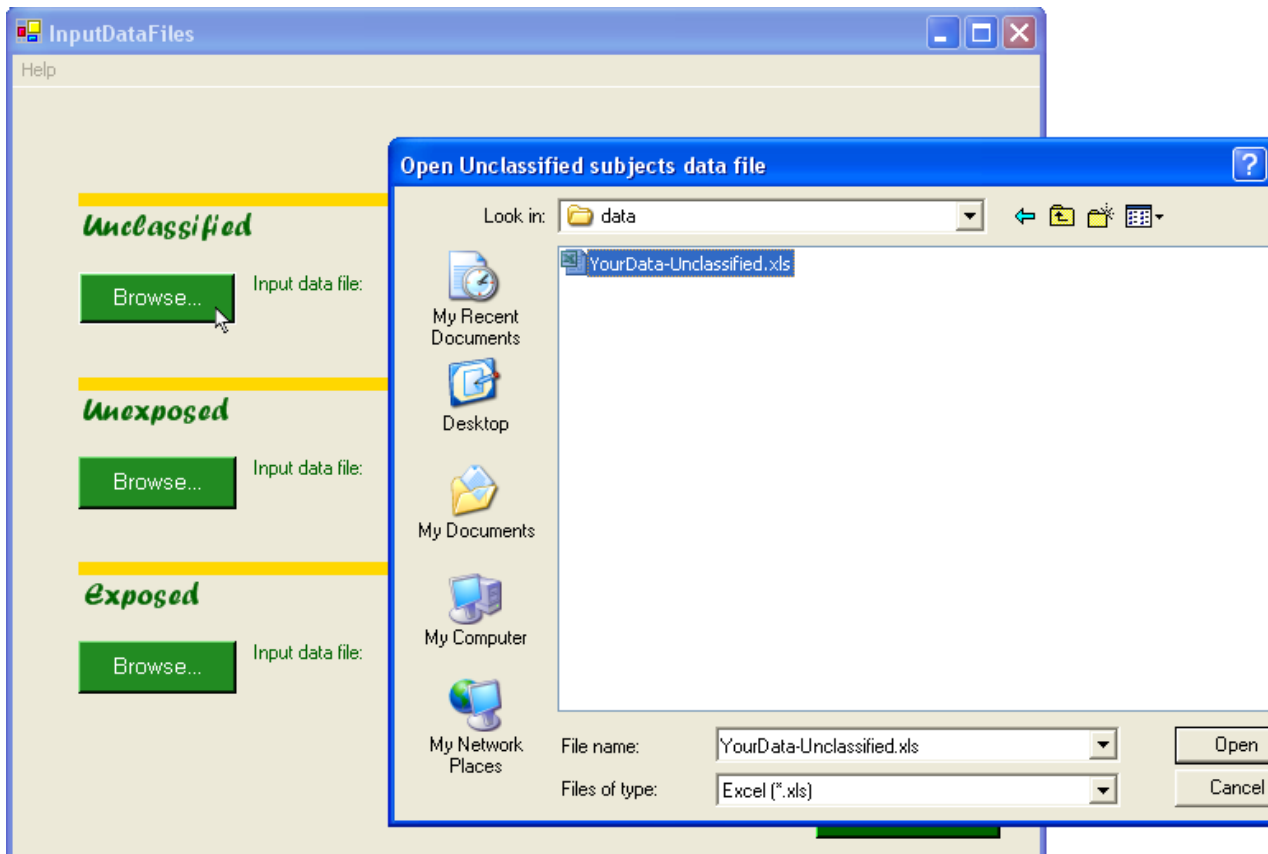
will be used where there is no ID variable defined. For example if Subject ID variables are defined in the exposed and unexposed files but not in the unclassified file, then the labels given in the first two files will be used, and the default label “Unclassified“ will be used for subjects in the unclassified file.

4. How to run ExposureAssessment

Three types of inputs are required for running this program:

- Comma-separated values (.csv) input data files (as just described in Section 3);
- Prior distributions for each unknown parameter (see Section 2);
- Initial values for each unknown parameter.

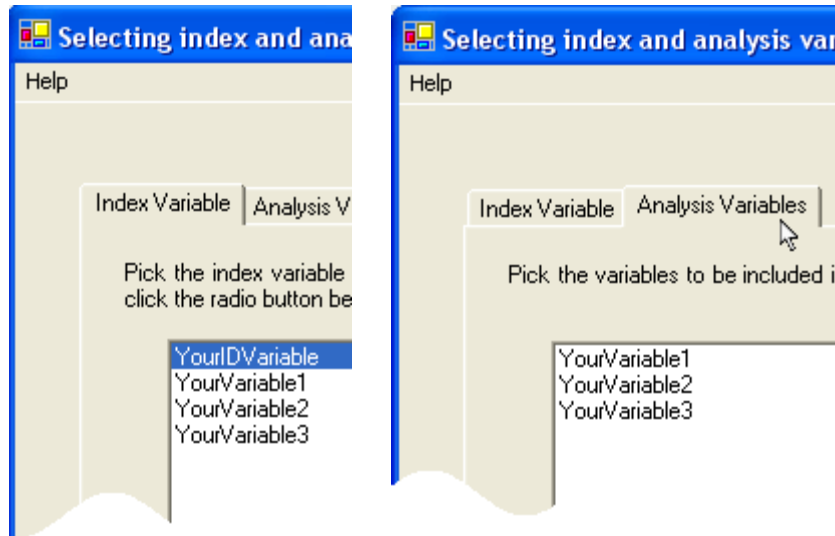
A first form (see below) with three Browse buttons allows the user to select each input data file in the appropriate (depending on whether the file consists in a list of Unclassified, Unexposed or Exposed subjects) section.



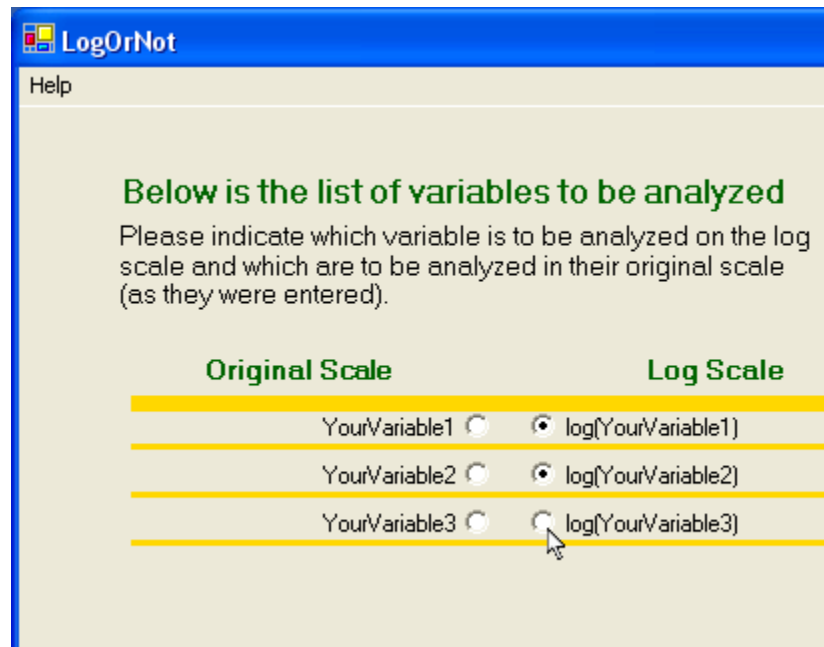
ExposureAssessment scans the first row of each input data file and reads in the different variables names (case insensitive).

These variables can include an Identity (or Index) variable and one or more analysis variables.

Not all variables found in input data files need to be included in the analysis.



The next form allows you to choose between the original scale and the log scale for each analysis variable: pick the log scale for a variable if you are going to model its log scores (which **ExposureAssessment** will compute) rather than the original scores.



The next form is used to enter your prior information on the prevalence of the disease, which is given a beta density with parameters (α, β) , such that prior mean and variance are $\alpha/(\alpha+\beta)$ and $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$, respectively.

The screenshot shows a window titled "Prevalence" with a "Help" button in the top left. The main content area contains the text "Please enter Prevalence prior parameters" in green. Below this, there are two input fields labeled α and β . To the right of these fields is a gold button with the text $(\alpha, \beta) \leftrightarrow (\mu, \sigma)$. Below the input fields is a label "Prior label" followed by a text input field.

$(\alpha, \beta) \leftrightarrow (\mu, \sigma)$

The gold button with text $(\alpha, \beta) \leftrightarrow (\mu, \sigma)$ allows you to specify your prior distributions in terms of prior moments (μ, σ) rather than in terms of (α, β) . If you choose to enter your prior information using (μ, σ) , the corresponding (α, β) values will be calculated automatically for you.

The next form (below) allows the user to fully describe the prior distributions for each unknown parameter used in the model, within both Exposed and Unexposed populations.

Please describe the prior distribution/knowledge each of the test scores used in this analysis in both the Exposed and Unexposed populations

Next >>

Exposed

Population mean: mean μ s.d. σ

Variation of individual means around population: lower limit upper limit

SD between:

Individual variation: lower limit upper limit

SD within:

Below-Detection Probability: α β

(α, β)
 (μ, σ)

Unexposed

Population mean: mean μ s.d. σ

Variation of individual means around population: lower limit upper limit

SD between:

Individual variation: lower limit upper limit

SD within:

Below-Detection Probability: α β

(α, β)
 (μ, σ)

Detection limit

Detection limit is: in original scale

This test doesn't have a detection limit.

Test label:

Next test description >

<< *log(yourvariable1)*

log(yourvariable2)

log(yourvariable3)

In the next two figures (right and below), we have enlarged the left and middle parts of figure above and superimposed (in gold) the corresponding variables names used in Section 2.

distribution on different tests used

Please describe the prior distribution/knowledge each of the U Unexposed populations

Exposed

Population mean mean μ s.d. σ

$\mu^{(1)}_1$ $\sigma^{2*(1)}_1$

Variation of individual means around population mean

lower limit upper limit

SD between $\sigma^{(1)}_{BL1}$ $\sigma^{(1)}_{BU1}$

Individual variation

lower limit upper limit

SD within $\sigma^{(1)}_{WLI}$ $\sigma^{(1)}_{WUI}$

Below-Detection Probability

α β

 $\alpha^{(1)}_{\epsilon 1}$ $\beta^{(1)}_{\epsilon 1}$

Detection limit

Detection limit is: ϵ_1 in original scale ▼

This test doesn't have a detection limit.

of the test scores used in this analysis in both the Exposed

Unexposed

Population mean mean μ s.d. σ

$\mu^{+(\theta)}_1$ $\sigma^{2+(\theta)}_1$

Variation of individual means around population

SD between lower limit upper limit

$\sigma^{(\theta)}_{BL1}$ $\sigma^{(\theta)}_{BUI}$

Individual variation

SD within lower limit upper limit

$\sigma^{(\theta)}_{WLI}$ $\sigma^{(\theta)}_{WUI}$

Below-Detection Probability

(α, β) α β

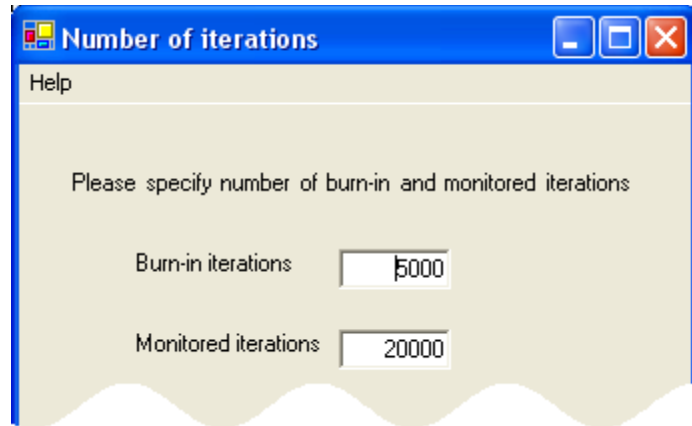
(μ, σ) $\alpha^{(\theta)}_{\epsilon 1}$ $\beta^{(\theta)}_{\epsilon 1}$

When the prior mean values entered into the Exposed and Unexposed population mean boxes differ, their values indicate whether higher or lower values in the corresponding test are expected in the Exposed population (see definition of δ_j in section 2).

When the prior means do not differ, the user will need to manually indicate whether higher or lower values in the corresponding test are expected in the Exposed population. A form will pop up in which the user must enter this information.

The Gibbs sampler specifications form (pictured at right) will allow you to control for the number of burn-in iterations and the number of monitored iterations.

Burn-in iterations are iterations that are ignored when the summary statistics are calculated and are used to allow the Markov chain to converge; the history plots in the .odc file (see Table 3 in Section 5) can be examined to assess convergence (of course, more formal convergence checks can be done, but this is beyond the scope of this document).



Number of iterations

Help

Please specify number of burn-in and monitored iterations

Burn-in iterations

Monitored iterations

You will then need to provide initial values for each unknown parameter, in both Exposed and Unexposed populations. These initial values are required in order to run any Gibbs sampler model, and need to be chosen carefully to ensure convergence to the proper posterior distributions. In general, you should pick values that are your best guesses as to the true values you expect for each parameter. You might also want to run the program several times with different starting values to ensure the Gibbs sampler converges to the same solution regardless of starting values.

initial values for each parameter below, for both Exposed and Unexposed .cls.

	<i>Exposed</i>	<i>Unexposed</i>
Population mean	<input type="text"/>	<input type="text"/>
SD between	<input type="text"/>	<input type="text"/>
SD within	<input type="text"/>	<input type="text"/>
Below-Detection Probability	<input type="text"/>	<input type="text"/>

Next test initial values >

*<< log(yourvariable1)
log(yourvariable2)
log(yourvariable3)*

Prevalence initial value

Next >>

Once all of the above required inputs are completed, **ExposureAssessment** will write a WinBUGS program and run it. Upon completion of the WinBUGS run, a form will pop up, allowing the user to view all output files produced.

5. An example of running **ExposureAssessment**

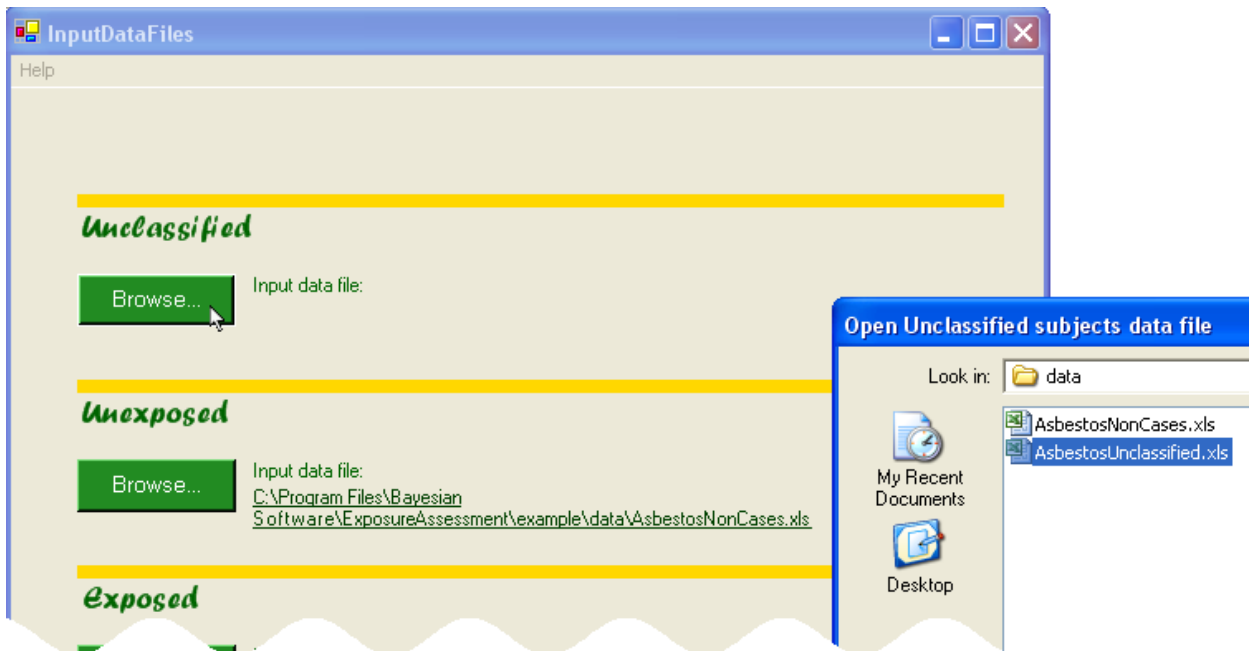
We will now illustrate the use of **ExposureAssessment** through an example, analyzing the same data as in the paper cited in introduction. We will use the same prior parameters as used in that paper, and the values are given in the table below.

		Exposed (g=1)	Unexposed (g=0)
Asbestos Bodies (on log scale)			
Population mean	$\mu^{(g)}_1$	$N(\mu^{*(1)}_1 = 8, \sigma^{2*(1)}_1 = 1)$	$N(\mu^{*(0)}_1 = 6, \sigma^{2*(0)}_1 = 1)$
SD Between	$\sigma^{(g)}_{B1}$	$U(\sigma^{(1)}_{BL1} = 0.01, \sigma^{(1)}_{BU1} = 3)$	$U(\sigma^{(0)}_{BL1} = 0.01, \sigma^{(0)}_{BU1} = 3)$
SD Within	$\sigma^{(g)}_{W1}$	$U(\sigma^{(1)}_{WL1} = 0.1, \sigma^{(1)}_{WU1} = 3)$	$U(\sigma^{(0)}_{WL1} = 0.1, \sigma^{(0)}_{WU1} = 3)$
Below-detection probability	$p^{(g)}_1$	$\text{Beta}(\alpha^{(1)}_{\epsilon 1} = 1, \beta^{(1)}_{\epsilon 1} = 1)$	$\text{Beta}(\alpha^{(0)}_{\epsilon 1} = 1, \beta^{(0)}_{\epsilon 1} = 1)$
Detection limit	ϵ_1	$\epsilon_1 = 40$	
Long Fibers (on log scale)			
Population mean	$\mu^{(g)}_2$	$N(\mu^{*(1)}_2 = 8, \sigma^{2*(1)}_2 = 1)$	$N(\mu^{*(0)}_2 = 5, \sigma^{2*(0)}_2 = 1)$
SD Between	$\sigma^{(g)}_{B2}$	$U(\sigma^{(1)}_{BL2} = 0.01, \sigma^{(1)}_{BU2} = 3)$	$U(\sigma^{(0)}_{BL2} = 0.01, \sigma^{(0)}_{BU2} = 3)$
SD Within	$\sigma^{(g)}_{W2}$	$U(\sigma^{(1)}_{WL2} = 0.1, \sigma^{(1)}_{WU2} = 3)$	$U(\sigma^{(0)}_{WL2} = 0.1, \sigma^{(0)}_{WU2} = 3)$
Below-detection probability	$p^{(g)}_2$	$\text{Beta}(\alpha^{(1)}_{\epsilon 2} = 1, \beta^{(1)}_{\epsilon 2} = 1)$	$\text{Beta}(\alpha^{(0)}_{\epsilon 2} = 1, \beta^{(0)}_{\epsilon 2} = 1)$
Detection limit	ϵ_2	$\epsilon_2 = 70$	
Short Fibers (on log scale)			
Population mean	$\mu^{(g)}_3$	$N(\mu^{*(1)}_3 = 8, \sigma^{2*(1)}_3 = 1)$	$N(\mu^{*(0)}_3 = 6, \sigma^{2*(0)}_3 = 1)$
SD Between	$\sigma^{(g)}_{B3}$	$U(\sigma^{(1)}_{BL3} = 0.01, \sigma^{(1)}_{BU3} = 3)$	$U(\sigma^{(0)}_{BL3} = 0.01, \sigma^{(0)}_{BU3} = 3)$
SD Within	$\sigma^{(g)}_{W3}$	$U(\sigma^{(1)}_{WL3} = 0.1, \sigma^{(1)}_{WU3} = 3)$	$U(\sigma^{(0)}_{WL3} = 0.1, \sigma^{(0)}_{WU3} = 3)$
Below-detection probability	$p^{(g)}_3$	$\text{Beta}(\alpha^{(1)}_{\epsilon 3} = 1, \beta^{(1)}_{\epsilon 3} = 1)$	$\text{Beta}(\alpha^{(0)}_{\epsilon 3} = 1, \beta^{(0)}_{\epsilon 3} = 1)$
Detection limit	ϵ_3	$\epsilon_3 = 70$	
Other			
Prevalence	π	$\text{Beta}(\alpha = 1, \beta = 1)$	

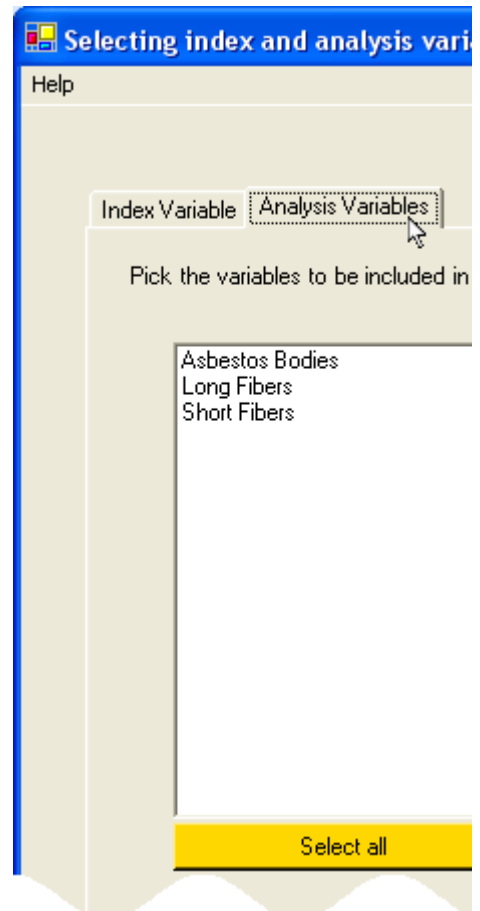
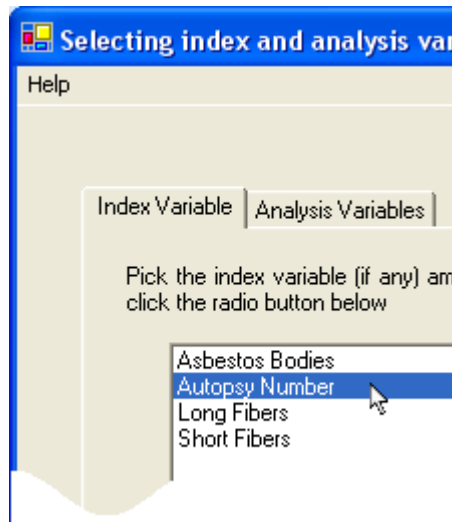
Table 1. Asbestos Fibers Exposure Assessment: Prior Distributions Hyperparameters.

Start **ExposureAssessment** by double-clicking the file **ExposureAssessment.vbs** (saved by default in c:\Program Files\Bayesian Software\ExposureAssessment or C:\Documents and Settings\user name\My Documents\Bayesian Software\ExposureAssessment, depending on your platform). There should also be a shortcut to **ExposureAssessment** in your Start menu.

Use the data files *AsbestosExposed.csv* and *AsbestosUnclassified.csv* (found in the example\data subdirectory of **ExposureAssessment**) as input files; as the file names indicate, *AsbestosExposed.csv* consists of data from (known) exposed subjects while *AsbestosUnclassified.csv* consists of data from a series of unclassified subjects for which we would like to compute the exposure probabilities.



Both input data files included the variables *Asbestos Bodies*, *Long Fibers* and *Short Fibers*, which will be the analysis variables. Autopsy Number was an identity number included in the file *AsbestosExposed.csv*; click it from the list so that it is used as subject label in the final outputs files.



Then click the Analysis Variables tab: pick the variables to include in the analysis from the list box, or click the *Select all* button below the list as a shortcut to include them all.

Select the Log Scale for each analysis variable, as the log scale happened to have been used for these data.

Click the *Next* button.

LogOrNot

Help

Below is the list of variables to be analyzed

Please indicate which variable is to be analyzed on the log scale and which are to be analyzed in their original scale (as they were entered).

Original Scale	Log Scale
Asbestos Bodies <input type="radio"/>	<input checked="" type="radio"/> log(Asbestos Bodies)
Long Fibers <input type="radio"/>	<input checked="" type="radio"/> log(Long Fibers)
Short Fibers <input type="radio"/>	<input checked="" type="radio"/> log(Short Fibers)

Since little prior information was available on the prevalence of the disease in this particular sample, indicate a uniform prior distribution ($\alpha=\beta=1$). You may want to enter a label for this non informative prior if you feel it is going to be used again in the future.

Click the *Next* button.

Prevalence

Help

Please enter Prevalence prior parameters

α β

Prior label

In the next form (below), enter the hyperparameter values for the first analysis variable (highlighted on the right-hand side of the form, ***log(asbestos bodies)*** in the present case). Enter a label in the *Test label* text box to make the use of the same prior description only one-click away the next time you run **ExposureAssessment**.

The screenshot shows a software window titled "Distribution on different tests used". The main instruction is: "Please describe the prior distribution/knowledge each of the test scores used in this analysis in both the Exposed and unexposed populations".

The interface is divided into two main sections: "Exposed" and "Unexposed".

Exposed Section:

- Population mean: mean μ = 8, s.d. σ = 1
- Variation of individual means around population: SD between lower limit 0.01, upper limit 3
- Individual variation: SD within lower limit 0.1, upper limit 3
- Below-Detection Probability: α = 1, β = 1

Unexposed Section:

- Population mean: mean μ = 6, s.d. σ = 1
- Variation of individual means around population: SD between lower limit 0.01, upper limit 3
- Individual variation: SD within lower limit 0.1, upper limit 3
- Below-Detection Probability: α = 1, β = 1

Other Fields:

- Detection limit: Detection limit is: 40 in original scale
- This test doesn't have a detection limit.
- Test label: Asb Bodies

Navigation and Labels:

- A green "Next >>" button is at the top right.
- A yellow bar highlights the label "<< ***log(asbestos bodies)***".
- Below it are two green labels: "***log(long fiber)***" and "***log(short fibers)***".
- A yellow "Next test description >" button is at the bottom center.

Click the *Next test description* button or the bold label ***log(long fibers)*** at the right of the form to proceed with the entry of *Long fibers* prior distribution parameters.

istribution on different tests used

se describe the prior distribution/knowledge each of the test scores used in this analysis in both the Exposed and Unexposed populations

Exposed

Population mean: mean μ s.d. σ

Variation of individual means around population mean: lower limit upper limit
SD between:

Individual variation: lower limit upper limit
SD within:

Below-Detection Probability: α β

Unexposed

Population mean: mean μ s.d. σ

Variation of individual means around population mean: lower limit upper limit
SD between:

Individual variation: lower limit upper limit
SD within:

Below-Detection Probability: α β

Detection limit: Detection limit is: in original scale
 This test doesn't have a detection limit.

Test label:

Priors previously used for log(long fibers)

-

Priors previously used for other tests

-

log(asbestos bodies)
<< log(long fibers)
log(short fiber)

Since we have run **ExposureAssessment** with both *Long fibers* and *Short fibers* before running the present example and have saved the prior parameters used by entering a label in the *Test label* text box, we can take advantage of the shortcuts to priors descriptions listed in bottom form list boxes (above): when clicking the appropriate prior label (*Lg fibers*, above), hyperparameters text boxes will be automatically filled with the values entered when that prior label was last used.

For the time being, however, you have never entered a prior for *Long fibers*: thus you will need to type in the values shown in figure above (also found in Table 1).

Proceed the same way for *Short Fibers* prior description and click the *Next* button to proceed to initial value entry. The initial values for each unknown parameter in this example are listed in Table 2.

		<i>Exposed</i> (g=1)	<i>Unexposed</i> (g=0)
<i>Asbestos Bodies (on log scale)</i>			
Population mean	$\mu^{(g)}_1$	8	6
SD Between	$\sigma^{(g)}_{B1}$	1.5	1
SD Within	$\sigma^{(g)}_{W1}$	0.8	0.6
Below-detection probability	$p^{(g)}_1$	0.05	0.5
<i>Long Fibers (on log scale)</i>			
Population mean	$\mu^{(g)}_2$	8	5
SD Between	$\sigma^{(g)}_{B2}$	1.5	1
SD Within	$\sigma^{(g)}_{W2}$	0.8	0.6
Below-detection probability	$p^{(g)}_2$	0.05	0.5
<i>Short Fibers (on log scale)</i>			
Population mean	$\mu^{(g)}_3$	8	6
SD Between	$\sigma^{(g)}_{B3}$	1.5	1
SD Within	$\sigma^{(g)}_{W3}$	0.8	0.6
Below-detection probability	$p^{(g)}_3$	0.05	0.3
<i>Other</i>			
Prevalence	π	0.5	

Table 2. Asbestos Fibers Exposure Assessment: Parameters Initial Values.

Initial values

Help

Enter initial values for each parameter below, for both Exposed and Unexposed subjects.

	<i>Exposed</i>	<i>Unexposed</i>
Population mean	<input type="text" value="8"/>	<input type="text" value="6"/>
SD between	<input type="text" value="1.5"/>	<input type="text" value="1"/>
SD within	<input type="text" value="0.8"/>	<input type="text" value="0.6"/>
Below-Detection Probability	<input type="text" value="0.05"/>	<input type="text" value="0.5"/>

Next test initial values >

Prevalence initial value

Next >>

<< log(asbestos bodies)
log(long fibers)
log(short fibers)

Enter initial values (see Table 2) for each unknown parameter for each analysis variable as well as for Prevalence, in the lower-right corner of the form. Click the *Next* button when done.

Problem Description Reviewal

This page summarizes all of the information you have entered. Please check all information carefully. If all is correct proceed to last form by clicking on the "Next" button. If you want to change any of the inputs you have provided, click on the incorrect value, which will bring you back to that parameters input screen.

Data files
Unexposed C:\patrick\DiagnosticTests\ExposureAssessment\pka\ex\data\AsbestosNonCases.xls
Unclassified C:\patrick\DiagnosticTests\ExposureAssessment\pka\ex\data\AsbestosUnclassified.xls

Prevalence **Gibbs sampling**
Prior distribution *Initial value* *Burn-in:* 5000
 Beta ($\alpha=1, \beta=1$) 0.5 *Monitored:* 20000

	Population mean		SD Between		SD Within		Below-Detection Probability	
	<i>Prior distribution</i>	<i>Initial value</i>	<i>Prior distribution</i>	<i>Initial value</i>	<i>Prior distribution</i>	<i>Initial value</i>	<i>Prior distribution</i>	<i>Initial value</i>
<i>log(asbestos bodies)</i>								
▶ Exposed	N(8, 1)	8	U(0.01, 3)	1.5	U(0.1, 3)	0.8	Beta ($\alpha=1, \beta=1$)	0.05
Unexposed	N(6, 1)	6	U(0.01, 3)	1	U(0.1, 3)	0.6	Beta ($\alpha=1, \beta=1$)	0.5
<i>log(long fibers)</i>								
▶ Exposed	N(8, 1)	8	U(0.01, 3)	1.5	U(0.1, 3)	0.8	Beta ($\alpha=1, \beta=1$)	0.05
Unexposed	N(5, 1)	5	U(0.01, 3)	1	U(0.1, 3)	0.6	Beta ($\alpha=1, \beta=1$)	0.5
<i>log(short fibers)</i>								
▶ Exposed	N(8, 1)	8	U(0.01, 3)	1.5	U(0.1, 3)	0.8	Beta ($\alpha=1, \beta=1$)	0.05
Unexposed	N(6, 1)	6	U(0.01, 3)	1	U(0.1, 3)	0.6	Beta ($\alpha=1, \beta=1$)	0.3

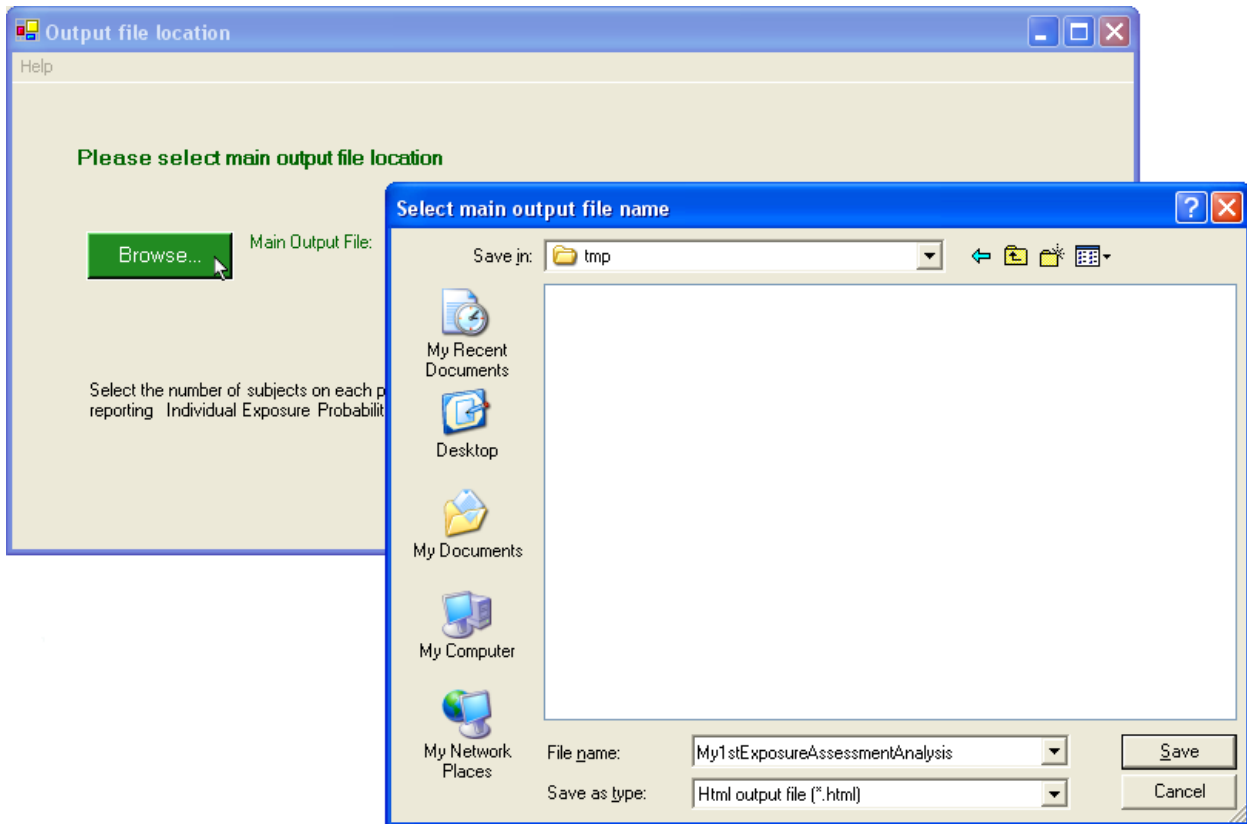
▶ Solid triangles indicate the group with higher expected scores

Next >>

In the above form, almost every bit of information entered so far can be checked and modified if necessary. Hovering over a modifiable unit of information (such as prior distribution hyperparameters or initial values) with the mouse will turn the pointer into a hand: clicking on that item will allow you to revisit the values entered in the corresponding original form: when necessary, modify the incorrect entry in the form and click *Back to Problem Reviewal* button.

When every piece of information has been verified as correct and you are ready to proceed to parameters estimation, click the *Next* button.

A final form will allow you to select the Main Output File location (an .html file). Note that secondary output files will also produced (click the *More* button for more information). If any of the Main or Secondary output files will overwrite an existing file, you will be prompted with a warning message to this effect.



On the same form (hidden here by the *Select main output file name* form), the user can select the number of subjects to be plotted on each page in reporting Individual Exposure Probabilities

Upon completion, **ExposureAssessment** will pop up a final form with contains links allowing you to view all main and secondary output files. The output files produced are listed in Table 3.

Main output files	
<output file name>.html	<p>Main ExposureAssessment output file.</p> <p>Contains:</p> <ul style="list-style-type: none"> • Individual Exposure Probabilities (median and 95% credible intervals); • Posterior distributions (medians and 95%) for each unknown parameter; • Number of burn-in and monitored iterations; • Prior distributions used; • Initial values used. <p>Also contains links to Secondary output files.</p>
<output file name>.odc	<p>WinBUGS odc output file (a complete binary file that can be opened in WinBUGS).</p> <p>This file is produced by WinBUGS, not by ExposureAssessment directly.</p>
Secondary output files	
<output file name>-data.html	<p>Html data presentation.</p> <p>Always good to have a look at that file to make sure the right data were analyzed.</p>
<output file name>.txt	<p>WinBUGS text output file.</p> <p>Also produced by WinBUGS; does not contain information not already contained in the .odc output file, but is somewhat easier to consult, if necessary (although the Main ExposureAssessment .html output file should already contain everything you need to know from this file).</p>
<output file name>.pdf	<p>Plot of Individual Exposure Probabilities with 95% credible intervals.</p>

Table 3. ExposureAssessment output files.

Of course, these output files will not be deleted when you close the final form; you will still be able to view these files by browsing to their location with the Windows Explorer.

The top part of the main html output file reports statistics on the Individual Exposure Probabilities. Subjects are sorted by descending median.

In the excerpt presented here, many subjects were almost certainly exposed, with a median posterior probability equal to 1. The first bunch even have a 95% Credible Interval lower limit equal to 1, which translates a high certainty about the diagnosis.

Other subjects, such as Unclassified-42, 32, 12, and 15 also had a high Estimated probability median: however, their 95% Credible Interval do not exclude very low exposure probabilities.

Following subjects (Unclassified-24, 33, 21,13 and 77) present a very low Estimated probability median (virtually 0) but also have a wide 95% Credible Interval.

The remaining subjects were almost surely unexposed, with Exposure Probabilities concentrated around 0.

Individual Exposure Probabilities

Autopsy Number	Estimated probability		95% Credible Interval	
	Median	Mean	lower limit	upper limit
Unclassified-2	1.0	1.0	1.0	1.0
Unclassified-3	1.0	1.0	1.0	1.0
Unclassified-5	1.0	1.0	1.0	1.0
Unclassified-6	1.0	1.0	1.0	1.0
Unclassified-7	1.0	1.0	1.0	1.0
Unclassified-10	1.0	1.0	1.0	1.0
Unclassified-16	1.0	1.0	1.0	1.0
Unclassified-17	1.0	1.0	1.0	1.0
Unclassified-22	1.0	1.0	1.0	1.0
Unclassified-23	1.0	1.0	1.0	1.0
Unclassified-25	1.0	1.0	1.0	1.0
Unclassified-26	1.0	1.0	1.0	1.0
Unclassified-27	1.0	1.0	1.0	1.0
Unclassified-28	1.0	1.0	1.0	1.0
Unclassified-73	1.0	1.0	1.0	1.0
Unclassified-79	1.0	1.0	1.0	1.0
Unclassified-14	1.0	0.9913	0.9937	1.0
Unclassified-42	1.0	0.8935	2.022E-28	1.0
Unclassified-32	1.0	0.8538	4.224E-14	1.0
Unclassified-12	0.9998	0.7611	3.138E-16	1.0
Unclassified-15	0.9985	0.586	1.319E-21	1.0
Unclassified-24	5.863E-7	0.3797	2.754E-31	1.0
Unclassified-33	1.206E-7	0.3473	4.199E-32	1.0
Unclassified-21	8.314E-9	0.4252	8.154E-40	1.0
Unclassified-13	2.681E-9	0.4031	2.782E-39	1.0
Unclassified-35	4.07E-10	3.938E-4	2.634E-24	7.203E-4
Unclassified-75	6.624E-11	6.181E-6	2.431E-23	1.169E-5
Unclassified-19	5.25E-11	0.002825	1.177E-29	0.004063
Unclassified-41	1.587E-11	6.05E-6	6.541E-24	6.249E-6
Unclassified-50	1.331E-12	2.004E-4	4.126E-32	2.428E-4
Unclassified-58	1.317E-12	4.317E-5	9.276E-31	4.355E-5
Unclassified-44	1.904E-13	0.01967	9.602E-36	0.09487
Unclassified-53	1.783E-13	5.829E-5	1.876E-32	2.116E-5
Unclassified-18	4.146E-14	4.854E-4	1.812E-34	1.355E-4
Unclassified-11	5.615E-15	1.156E-6	6.713E-31	8.212E-7
Unclassified-71	2.782E-16	1.583E-5	6.765E-39	4.697E-6
Unclassified-77	1.713E-16	0.339	0.0	1.0
Unclassified-9	1.041E-16	0.002863	1.241E-39	7.324E-5
Unclassified-65	7.468E-17	2.077E-7	3.296E-34	5.543E-8
Unclassified-34	1.3E-17	2.054E-8	2.599E-33	4.631E-9
Unclassified-57	1.111E-18	0.1133	0.0	1.0
Unclassified-76	4.214E-19	1.169E-5	2.08E-41	1.963E-7
Unclassified-4	3.356E-19	4.669E-9	1.369E-36	6.547E-10
Unclassified-78	2.441E-19	2.044E-10	1.622E-40	3.89E-10
Unclassified-55	1.621E-21	8.818E-11	2.537E-40	1.744E-11
Unclassified-20	4.909E-22	1.857E-8	4.204E-45	8.504E-10
Unclassified-74	1.095E-23	4.839E-11	1.447E-41	1.106E-12
Unclassified-56	5.231E-24	6.695E-10	0.0	4.868E-10
Unclassified-62	8.444E-25	0.1703	0.0	1.0
Unclassified-8	6.706E-27	4.033E-7	0.0	6.22E-11
Unclassified-66	1.96E-28	1.932E-6	0.0	3.512E-10
Unclassified-46	1.41E-30	2.191E-8	0.0	2.469E-13

The next part of the main output file summarizes posterior distribution for each exposure variable.

In this example, the Below-Detection Probabilities ($p_j^{(g)}$, $j=1,2,3$, $g=1,2$, see Section 2) show much higher probabilities of undetectable values in the Unexposed Population than in the Exposed population, which goes with intuition.

Posterior means are higher in the Exposed population than in the Unexposed population for each exposure variable by a fair margin (from 2.5 to 3 points). The non-crossing 95% credible intervals for means shows that the data allowed a clear distinction between Exposed and Unexposed subjects.

Posterior distributions

Exposed and Unexposed posterior distributions description

Variable		Below-Detection Probability			Mean		
		Median	95% credible interval limits		Median	95% credible interval limits	
			lower	upper		lower	upper
log(Asbestos Bodies)	Exposed	0.05909	0.02858	0.1027	8.859	8.437	9.309
	Unexposed	0.4387	0.3524	0.5354	5.889	5.444	6.29
log(Long Fibers)	Exposed	0.03216	0.0112	0.0674	7.973	7.602	8.37
	Unexposed	0.5422	0.4511	0.6465	5.452	5.274	5.622
log(Short Fibers)	Exposed	0.02939	0.0105	0.06276	8.996	8.624	9.368
	Unexposed	0.2975	0.2243	0.3851	6.256	5.98	6.511

The rightmost part of the Posterior distributions (below) summarizes the posterior distributions for both Between-subjects SD ($\sigma_{Bj}^{(g)}$, $j=1,2,3$, $g=0,1$, Section 2) and Within-subjects SD ($\sigma_{Wj}^{(g)}$, $j=1,2,3$, $g=0,1$). Between-subjects variation is somewhat larger in the Exposed population than in the Unexposed population for each exposure variable, as one would expect. Within-subjects is not negligible, as can be seen from reported posterior medians and/or 95% credible intervals.

Posterior distributions

Exposed and Unexposed posterior distributions description

Variable		Below-Detection Probability	SD Between			SD Within		
			Median	95% credible interval limits		Median	95% credible interval limits	
			Median	lower	upper	Median	lower	upper
log(Asbestos Bodies)	Exposed	0.05909	1.445	1.182	1.778	0.9296	0.816	1.074
	Unexposed	0.4387	1.057	0.7694	1.405	0.9096	0.7364	1.158
log(Long Fibers)	Exposed	0.03216	1.18	0.9459	1.465	0.866	0.7616	0.9975
	Unexposed	0.5422	0.3309	0.1187	0.5134	0.4523	0.3425	0.6063
log(Short Fibers)	Exposed	0.02939	1.115	0.891	1.404	0.8011	0.7044	0.9167
	Unexposed	0.2975	0.4928	0.1043	0.7697	0.6155	0.4892	0.7379

See original article for full details.

6. Monte Carlo Markov Chain (MCMC) can be dangerous

This section aims to introduce the novice to basic MCMC ideas and to the potential traps to avoid in order to obtain valid results when using MCMC in general, and **ExposureAssessment** in particular. This is a very brief overview. Please consult a textbook such as *Markov chain Monte Carlo in practice*, Walter R. Gilks, Sylvia Richardson, and David. J. Spiegelhalter, Chapman and Hall, 1995.

Let $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ be the complete set of unknown parameters in a problem; in the problem addressed by **ExposureAssessment**, these parameters would be the different scales normal means, SD Between, SD Within and Below-Detection probabilities in both Exposed and Unexposed populations (as well as a set of latent disease status for each unclassified subject).

A Gibbs sampler algorithm proceeds as follows: given a set of initial values for each parameter, it samples a value for θ_1 from its conditional distribution; it samples a value for θ_1 that seems likely given the data AND the other parameters, temporarily considered as fixed. It then proceeds with second parameter (θ_2) and samples a value from its conditional distribution given the data and other parameters ($\theta_1, \theta_3, \theta_4, \dots, \theta_p$), and so on. Once each of the p parameters were sampled (from their respective conditional distributions), it starts again with θ_1 and repeats the process for a second iteration.

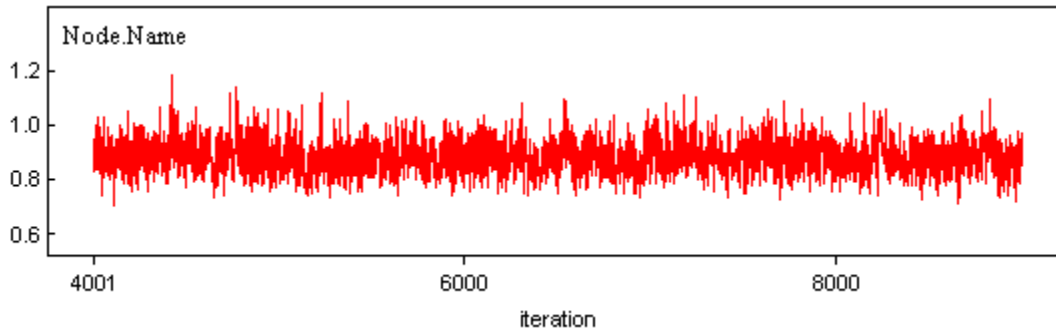
In the long run, the values sampled for a given parameter represent a sample from its marginal posterior distribution, which is the distribution of interest.

The values obtained at each iteration for a same parameter can be plotted in a *time-series-like* plot: on the y-axis we plot the value taken by the parameter at iteration i vs the iteration number on the x-axis, for each iteration, leading to a plot that is called the **trace** of that parameter.

Once the results from a Monte Carlo Markov Chain model are obtained, one should always remember to look at the different parameters traces and posterior distributions in order to assess the behaviour of the algorithm and to validate the appropriateness of the prior distributions and initial values used. Of course, these prior distributions and initial values should be carefully chosen in the first place to make sense clinically, but even with this preventive careful thinking, problems at the simulation step of project are not impossible, hence the importance of the following additional post-simulation checks.

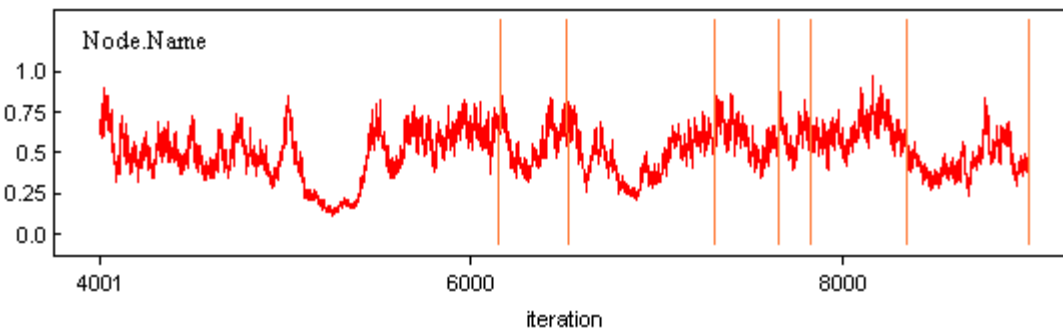
Ideally, the possible range of values (or the domain) of a parameter should be visited equally likely at any point in time, that is, the sampled values should not be restricted to a confined area for some time: that would depict an auto-correlation between successive estimates for that parameter. Even though it is sometimes (very) difficult to avoid in complex models, auto-correlation should be avoided as much as possible.

The trace below illustrates the ideal scenario:



Indeed, the algorithm seems to visit the range of likely values (from 0.8 to 1.0, roughly, in this example) for that variable (or node, as called in WinBUGS) in a very reasonable way, that is, high or low values seem to be visited equally likely at any point in the random walk.

The next trace shows an example where that goal is not really reached (ignore the orange vertical bars for the moment).



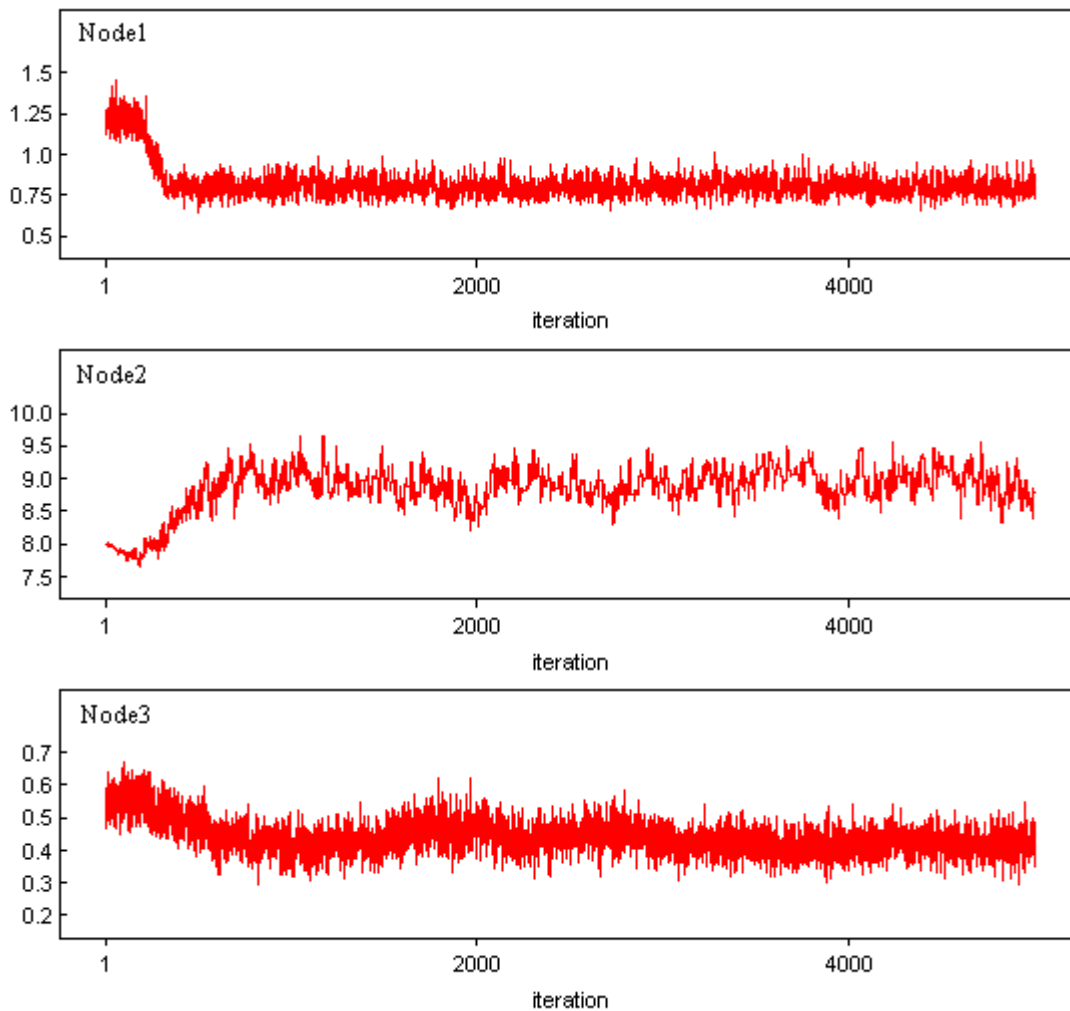
Indeed, after visiting the likely larger possible values, e.g., shortly after iteration 8000 (see the peak to the right of iteration 8000), it looks like the random walk leads to generally decreasing values, reaches a bottom limit, and then goes up for another while. This is obvious auto-correlation and in an ideal world should be avoided. However, in a package where you do not have control over the way the MCMC algorithm is run, you do not have much choice but to accept it. In addition, the most important is not really to avoid that auto-correlation phenomenon, but to take it into account when running your final simulation to ensure a sufficiently large number of cycles. A plot such as the one above indicates that there may be a problem, and running a larger number of iterations is advised. Even this is not sufficient to guarantee convergence, and at this point, the novice may wish to consult with a statistician experienced with MCMC convergence issues.

A cycle is a series of iterations where the algorithm seems to have visited the range of possible values for a variable. The cycles in the second part of the random walk traced above are roughly separated by vertical bars on the second half of the trace. In that part, there are roughly 6 cycles, that is, the algorithm has gone over the possible values at least 6 times.

The final number of iterations chosen in a WinBUGS run should ensure that a large number (hundreds or, even better, thousands) of cycles were performed for each node.

The number of burn-in iterations — that is, iterations that are dropped from inclusion when calculating the final inferences — is chosen to make sure that the algorithm has converged when monitoring of sampled values starts.

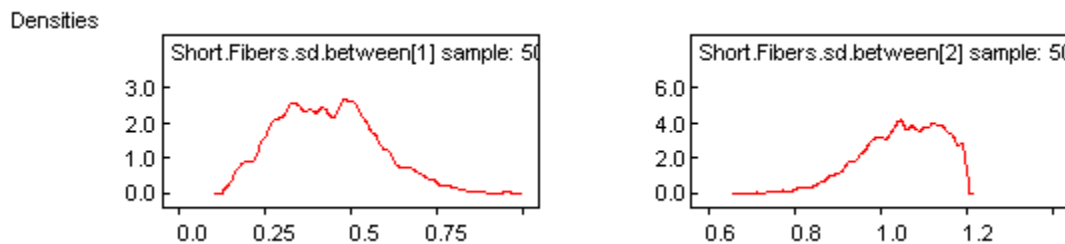
The traces below show the values taken for three nodes from the very first iteration. In each of them, it is clear that the first several hundred values taken by each of them differ from the rightmost more stable values. The trace plot for Node1 shows that the first 300 or so iterations are not in the same ballpark as the remainder, while the first 500 iterations for Node2 seem different from the rest. For Node3, it appears that a larger number of initial values is different from the rest, at least 1000 iterations, maybe 2000 iterations should be dropped. Consequently, in that problem, one should rerun the program with a burn-in of at least 2000 iterations.



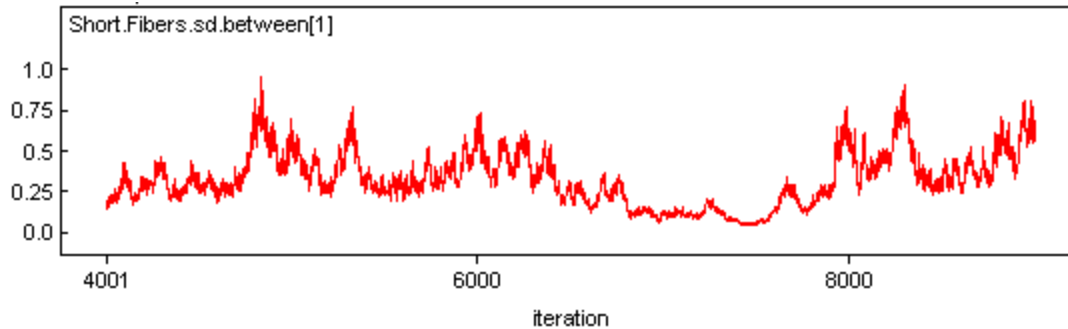
In any case, given the complexity of the problem addressed by **ExposureAssessment**, you cannot run the program with less than 4000 burn-in iterations. If the size of the data analyzed is reasonable and your computer fast enough, burn-in iterations should be cheap (in terms of running time): in that case, do not hesitate to burn-in even more iterations (perhaps 5, 10 or even 20 thousand).

In the problem addressed by **ExposureAssessment**, our experience leads us to believe that an informed choice of prior distributions and initial values usually leads to a good-mixing MCMC run, but this is never guaranteed for any specific data set, so that care is always needed.

Unless your prior distributions were based on very solid and uncontroversial scientific evidence, it is good practice to choose prior distributions that will let the data speak for themselves, that is, prior distributions that contain much less information about the prior parameter values than the information in the data themselves. For example, in the problem addressed by **ExposureAssessment**, the uniform distributions used on both Between and Within SDs should not be too narrow. In the example illustrated below, a uniform prior $U(0, 1.2)$ was used on Short Fibers' Between SD in exposed group: looking at its posterior density (node Short.Fibers.sd.between[2]), it is easy to realize that it is leaning towards its higher allowed values and that larger values may have been *appreciated* by the sampling algorithm, had larger values been allowed. Indeed, the posterior density falls down to 0 at its upper limit quite dramatically, showing a possibly too narrow prior being used.



The lower limit used on the uniform prior distributions for the SD parameters may also be problematic, although the problem is less likely to appear when many variables are used in the model. In the traces below, excerpted from an output where a uniform $U(0, 2)$ prior density was used for the Short Fibers Between SD parameter in the unexposed group, it seems like the sampling algorithm sometimes becomes stuck at very small values for the SD parameter, roughly between iterations 7000 to 7800.



This is clearly an undesired sampling phenomenon, for which a natural work-around may be to use a somewhat increased lower limit in the uniform prior for the problematic SD term, for example, something like $U(0.1, 2)$. A very small SD parameter is not likely in any case, as this implies that there is very little between-subject variability, which we know, in this particular problem, is highly implausible.

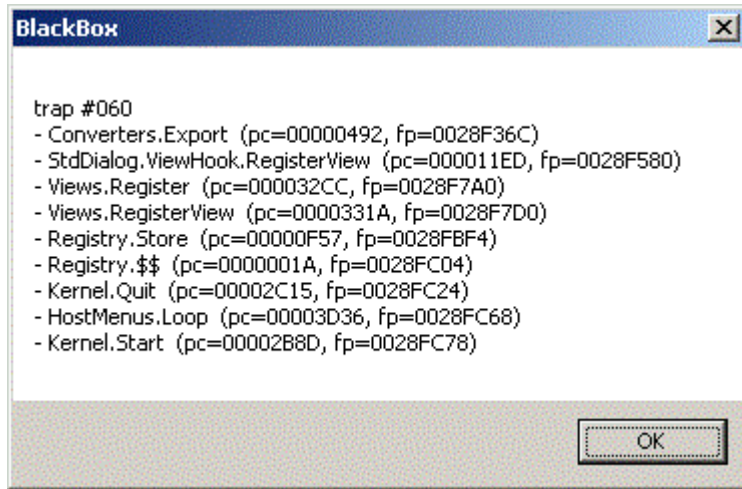
6.1 Sensitivity analysis

Finally, a sensitivity analysis should also be performed, that is, **ExposureAssessment** should be run a few times, each time with different (but still meaningful) prior distributions and/or initial values. The general idea is to check the changes in posterior inferences across a reasonable range of prior distributions. If the conclusions derived from each run are similar, then the conclusions can be considered as robust. If not, then the choice of prior distributions and the impact of prior choice on the conclusions needs to be more carefully assessed, and/or conclusions need to be drawn with some questions as to their robustness.

7.0 Avoiding Trap Errors on Windows 7 and Windows Vista platforms

If you are working on a Windows 7 or Windows Vista platform and have run WinBUGS before, you may have already run into the cryptic **Trap #060** error message illustrated to the right. This is due to restricted write permissions in c:\Program Files, where you may have installed WinBUGS.

WinBUGS **must** be installed in a directory where you have write permissions (e.g. C:\Users\user name \Documents) for **ExposureAssessment** to run smoothly.



8.0 Change log

Version 1.1 (July 2011)

Earlier versions used Excel input data files but did not work with Excel 2010 files. Hence the change to easier-to-read (programmatically, that is) Comma-Separated Values (.csv) input files.

Versions 1.2 and 1.2.1 (December 2011)

The previous default application folder (c:\Program Files) caused write permission problems for some Windows 7 and Vista users. Default application folder now changed to C:\Users\user name\Documents.

Versions 1.3 and 1.3.1 (February 2012)

Minor technical problem solved from previous version.

Versions 1.4 and 1.4.1 (April 2012)

We suggest a solution to prevent Trap errors for Windows 7 and Windows Vista.

Questions? Comments? Please send email to: lawrence.joseph@mcgill.ca

Other Bayesian software packages are available at
<http://www.medicine.mcgill.ca/epidemiology/Joseph>