Sample Size Calculations for Odds Ratio in presence of misclassification

(SSCOR Version 1.8.1, February 2019)

1. Introduction

The program SSCOR — available for Windows only — calculates sample size requirements for estimating odds ratio in the presence of misclassification. It is an implementation of the methods described in the paper

Bayesian sample size determination for case-control studies when exposure may be misclassified
Joseph L, Bélisle P
American Journal of Epidemiology. 2013;178(11):1673-1679.

We recommend that you carefully read the paper cited above before using this software.
You are free to use this program, for non-commercial purposes only, under two conditions:

- This note is not to be removed;
- Publications using **SSCOR** results should reference the manuscript mentioned above.

Please read the Install Instructions (InstallInstructions.html) prior to installation.

The easiest way to start SSCOR is to use the shortcut found in Programs list from the Start menu[1]. You will be prompted by a graphical user interface (GUI) to describe the required inputs, which include:

- choose between sample size calculations or outcome estimation for one or more sample sizes
- fill in your prior information about the prevalence of exposure in both case and control populations
- fill in your prior information about the probability of correct classification when the true exposure is positive or negative within both case and control populations
- (optional) attach labels to the prior distributions used; doing so will make the use of the same priors only one click away the next time you run **SSCOR**
- choose a sample size criterion (ACC, ALC or MWOC)
- indicate the location for your output file (where you want the results to be saved)
- (optional) answer a few more technical questions (number of Gibbs iterations, starting sample size, etc.). If you are unsure, the default values are well chosen for most common situations

Once you have input the above information, the program will search for the optimal sample size. In doing so, SSCOR will run a series of WinBUGS programs in a window you can minimize. After each WinBUGS run a C program will open to compute HPD intervals from the WinBUGS output, which will cause a MS-DOS window to pop-up. You can carry out other work while this is going on, and can ignore what is happening in the background. Running times can vary and could be several hours to even several days, depending on the required sample size, number of iterations within each WinBUGS program, and so on.

If you are calculating a single sample size, when the program has finished a popup window will appear giving you the opportunity to view the output immediately. This window will not appear when **SSCOR** is used to run a series of sample size estimations (from a series of *scripts*, see sections 3.1 and 3.1.1 below).

Each time **SSCOR** is run, a log file is saved under log\SSCOR.txt in the SSCOR home directory (C:\Users\user name\Documents\Bayesian Software\SSCOR or C:\Documents and Settings\user name\My Documents\Bayesian Software\ SSCOR, by default, depending on your platform). This log file is overwritten at each run.

You can refer to log file to retrieve error messages or confirm program success.

2. Estimating odds ratio in presence of misclassification

Consider a retrospective study in which a sample of known cases is obtained who have the outcome disease or characteristic of interest (D_1) and who are to be compared with an independent sample of non-diseased controls (D_0). For each subject in the two groups, the prior degree of exposure to the risk factor under study, classified as E_1 and E_0 for exposed and non-exposed, respectively, is then determined retrospectively, possibly with some degree of misclassification. While misclassification is often ignored, it can have a huge impact on odds ratio estimates. Consequently, sample size calculations can also be affected by misclassification.

When there is no misclassification error, the entries in a 2 x 2 table of frequencies are

	Cases (D_1)	Controls (D_0)	
E_1	a	b	
E_0	c	d	
	n_1	n_0	N

for fixed sample sizes n_1 and n_0.

Given retrospective samples of n_1 cases (D_1) and n_0 controls (D_0), the assumed conditional probabilities are

	Cases (D_1)	Controls (D_0)
E_1	θ_1	θ_0
E_0	$1-\theta_1$	$1-\theta_0$
	1.0	1.0

where θ_0 and θ_1 are probabilities of exposure conditional on disease status.

The retrospective odds ratio is given by

$$\varphi = \frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)} .$$

However, SSCOR does not address the case where exposure (E) is known exactly but rather measured through an imperfect surrogate E*, with possibly different sensitivities and specificities in the controls and cases groups, that is, with

$P\{E^* = 1 \mid E_1 \text{ (truly exposed) in group } D_i\} = s_i, \quad i = 0, 1.$
$P\{E^* = 0 \mid E_0 \text{ (truly unexposed) in group } D_i\} = c_i$

We thus observe the cell counts

	Cases (D_1)	Controls (D_0)	
$E^* = 1$	X_1	X_0	
$E^* = 0$	n_1-X_1	n_0-X_0	
	n_1	n_0	N

with disease conditional probabilities of apparent (that is, measured by an imperfect surrogate) exposure

	Cases (D_1)	Controls (D_0)
$E_* = 1$	θ^*_1	θ^*_0
$E_* = 0$	$1-\theta^*_1$	$1-\theta^*_0$
	1.0	1.0

where $\theta^*_i = \theta_i s_i + (1-\theta_i)(1-c_i)$, $i=0, 1$.

2.1 Model

Given sample sizes n_0 and n_1, the likelihood function is the product of two binomial distributions, since

$$X_i \sim \text{Binomial}(n_i, \theta*_i), \quad i=0, 1$$

where $\theta*_i = \theta_i s_i + (1-\theta_i)(1-c_i)$, $i = 0, 1$.

The prevalence of (true) exposure θ_i in both cases (i=1) and controls (i=0) is given a prior beta distribution

$$\theta_i \sim \text{Beta}(\alpha^{\theta}_{i,}\ \beta^{\theta}_{i}), \quad i=0, 1,$$

as well as the surrogate measure for exposure,

$$s_i \sim \text{Beta}(\alpha^{s}_{i,}\ \beta^{s}_{i})$$
$$c_i \sim \text{Beta}(\alpha^{c}_{i,}\ \beta^{c}_{i}), \quad i=0, 1.$$

The latter two can be made as close as necessary to a perfect surrogate (virtually without misclassification) when necessary (e.g., to compare sample size results obtained with low or moderate misclassification error to those obtained if there were no misclassification error). For example, using a beta(999999, 1) prior density is, for all practical purposes, equivalent to assuming no misclassification.

2.2 Stopping criterion

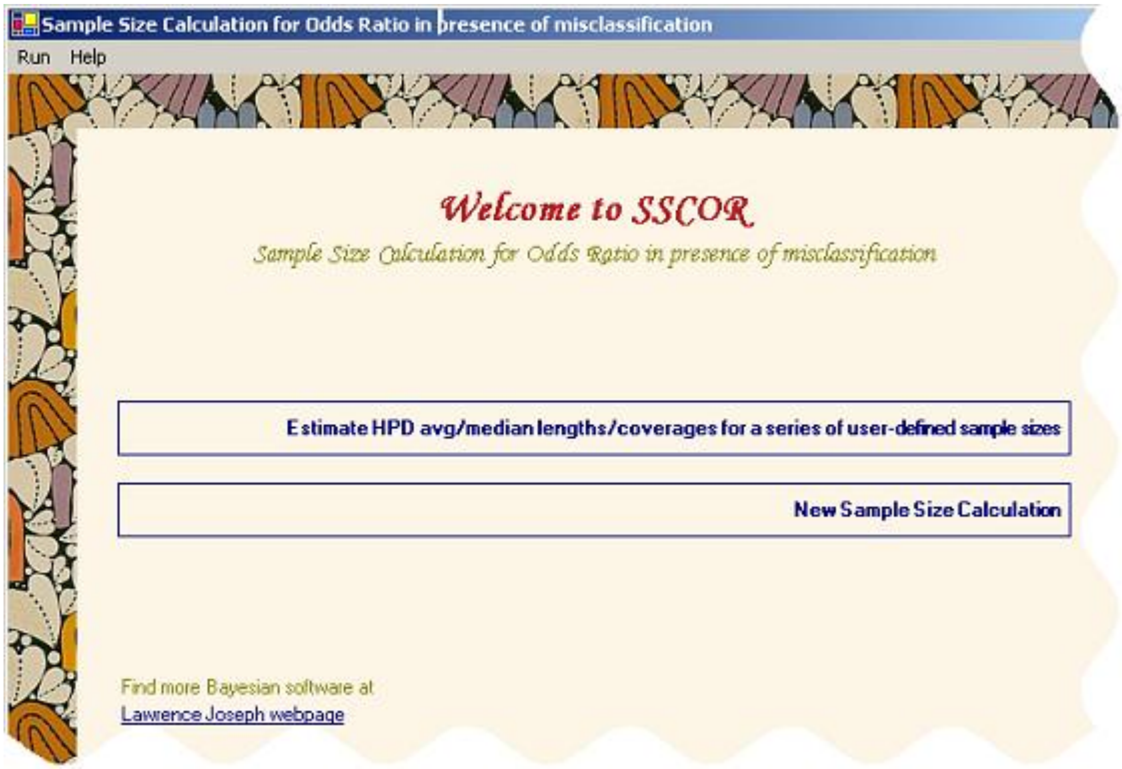
SSCOR iterates over *N* until

a) the desired parameter accuracy is met for sample size *N* but not for *N-1* or
b) in a series of six consecutive sample sizes, the larger three satisfy the sample size criterion while the smaller three do not, and these six consecutive sample sizes do not span more than 2% of their midpoint value.

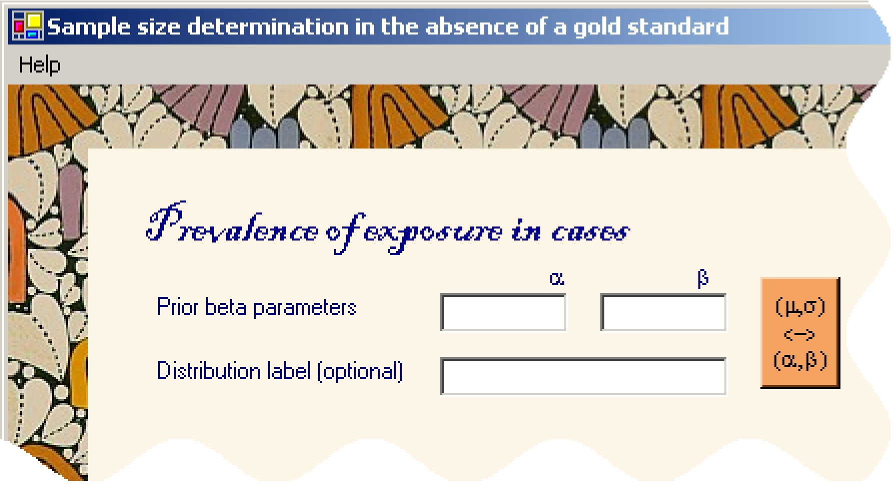
Stopping criterion (b) proves useful when the final sample size is large (e.g. more than a thousand).

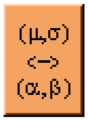
3. How to run SSCOR

Upon opening the program, the initial form allows you to indicate whether you are using the program to do actual sample size calculations, or to estimate HPD interval characteristics for a series of predetermined sample sizes.


Sample Size Calculation for Odds Ratio in presence of misclassification
Run
Help
Welcome to SSCOR
Sample Size Calculation for Odds Ratio in presence of misclassification
Estimate HPD avg/median lengths/coverages for a series of user-defined sample sizes
New Sample Size Calculation
Find more Bayesian software at
Lawrence Joseph webpage

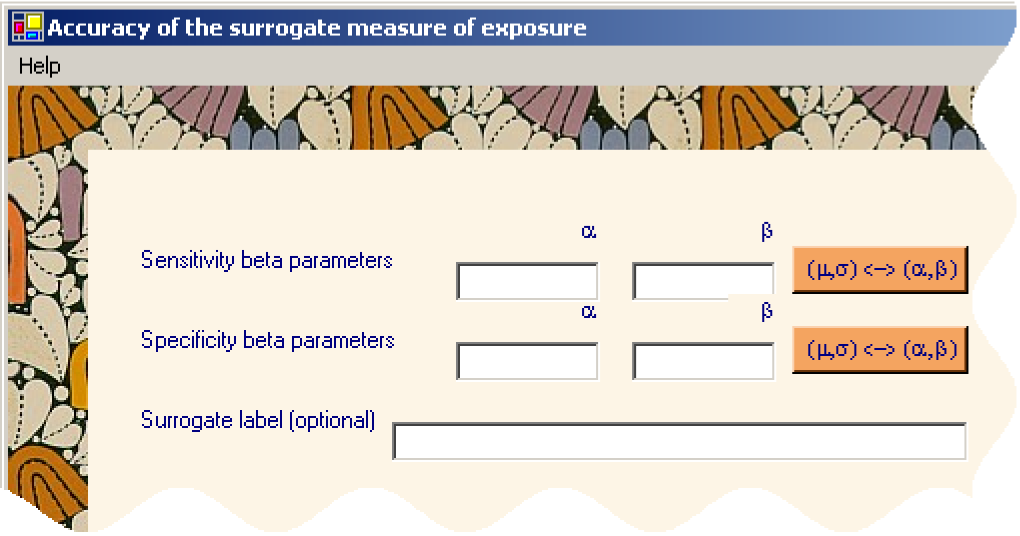
The next two forms are used to enter your prior information about the prevalence of exposure in cases and controls, in turn. Each is given a beta density with parameters (α,β), such that prior mean and variance are α/(α+β) and αβ/(α+β)^2(α+β+1), respectively.


Sample size determination in the absence of a gold standard
Help
Prevalence of exposure in cases
α
β
Prior beta parameters
Distribution label (optional)
(μ,σ)
<->
(α,β)



The orange button with text (μ,σ) <–> (α,β) allows you to specify your prior distributions in terms of prior moments (μ,σ) rather than in terms of beta parameters (α,β). If you choose to enter your prior information using (μ,σ), the corresponding (α,β) values will be calculated automatically for you.

We assume that exposure is measured with some degree of misclassification. The next form (pictured at right) allows one to enter the beta prior distribution parameters for both the sensitivity and the specificity of the surrogate for exposure.

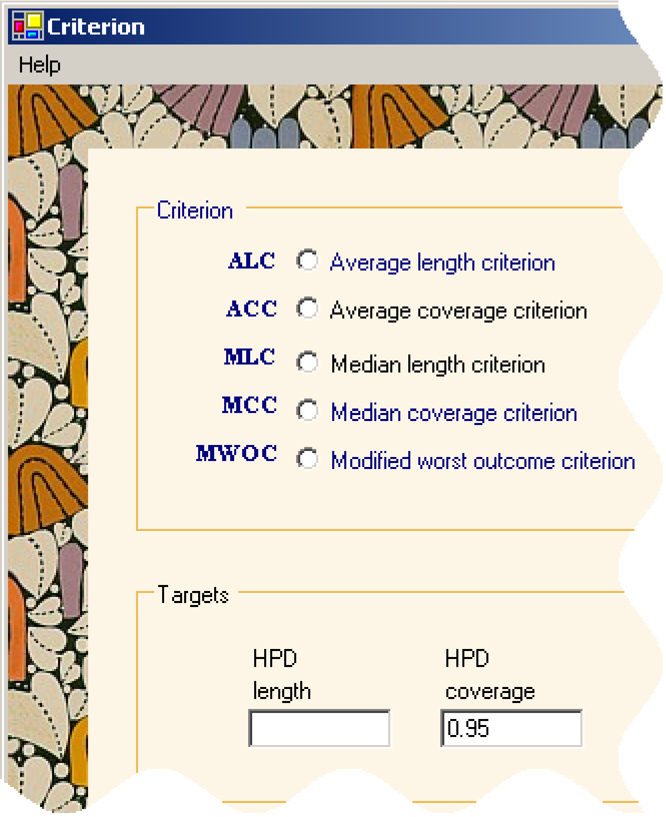

Accuracy of the surrogate measure of exposure
Help
α
β
Sensitivity beta parameters
(μ,σ) <-> (α,β)
α
β
Specificity beta parameters
(μ,σ) <-> (α,β)
Surrogate label (optional)

When running sample size calculations, the next step is the selection of a sample size criterion.

The form pictured right allows the user to pick the criterion on which to base sample size calculation, either ALC, ACC or MWOC.

Depending on the selected criterion, user will also be asked the HPD average or fixed length, HPD fixed or average coverage, as well as the MWOC-level when MWOC criterion is selected.

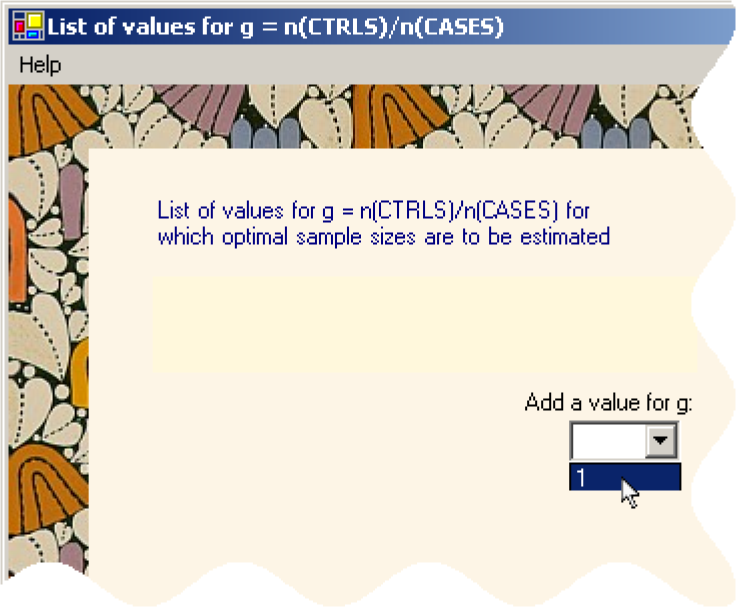
For a description of all of these criteria, please see the paper referenced at the beginning of this document.


Criterion
Help
Criterion
ALC ○ Average length criterion
ACC ○ Average coverage criterion
MLC ○ Median length criterion
MCC ○ Median coverage criterion
MWOC ○ Modified worst outcome criterion
Targets
HPD length

HPD coverage: 0.95

The ratio of the number of controls sampled to the number of cases sampled may depend on several factors, such as sampling costs that may be different between cases and controls, or availability of cases and controls in the population under study.

The next form (pictured at right) allows the user to select one or more values for this ratio on which to calculate sample size requirements. In general, each value of *g* will lead to different optimal sample sizes.


List of values for g = n(CTRLS)/n(CASES)

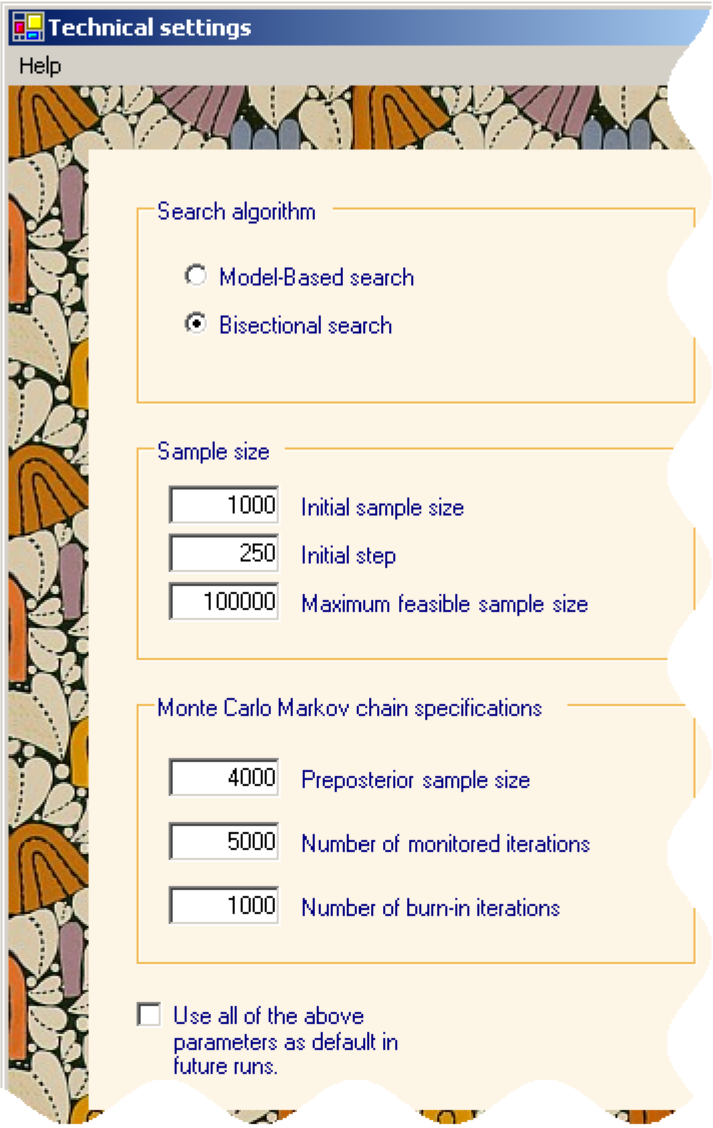
List of values for g = n(CTRLS)/n(CASES) for which optimal sample sizes are to be estimated

Add a value for g:

The next form allows the user to select the number of monitored iterations run for each WinBUGS program run, the number of burn-in iterations, the number of samples randomly selected at each sample size along the search (the preposterior sample size), as well as the initial sample size, the initial step to use in the search and the maximum feasible sample size (a sample size above which **SSCOR** will not go).

The top box allows the user to select whether the search for the sample size will be done through a bisectional search or a model-based search. The model based search will most often converge to the correct sample size in fewer steps (see Appendix A for details).

Once the sample size calculations are completed, **SSCOR** produces a scatter plot (one for each value of *g* chosen in previous form) of the outcome of interest (either HPD average length, average coverage, or some percentile of HPD coverages) versus each sample size visited along the search for optimal sample size. That scatter plot may help the user judge whether the model-based search was a good idea or not given his particular problem.



Search algorithm

○ Model-Based search

● Bisectional search

Sample size

1000 Initial sample size

250 Initial step

100000 Maximum feasible sample size

Monte Carlo Markov chain specifications

4000 Preposterior sample size

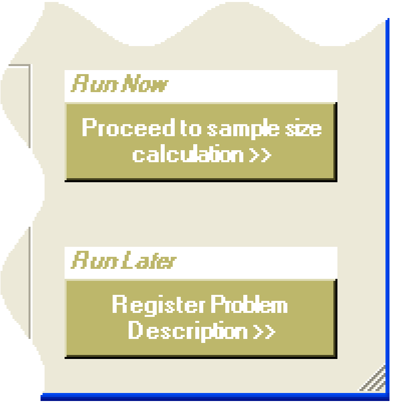
5000 Number of monitored iterations

1000 Number of burn-in iterations

□ Use all of the above parameters as default in future runs.

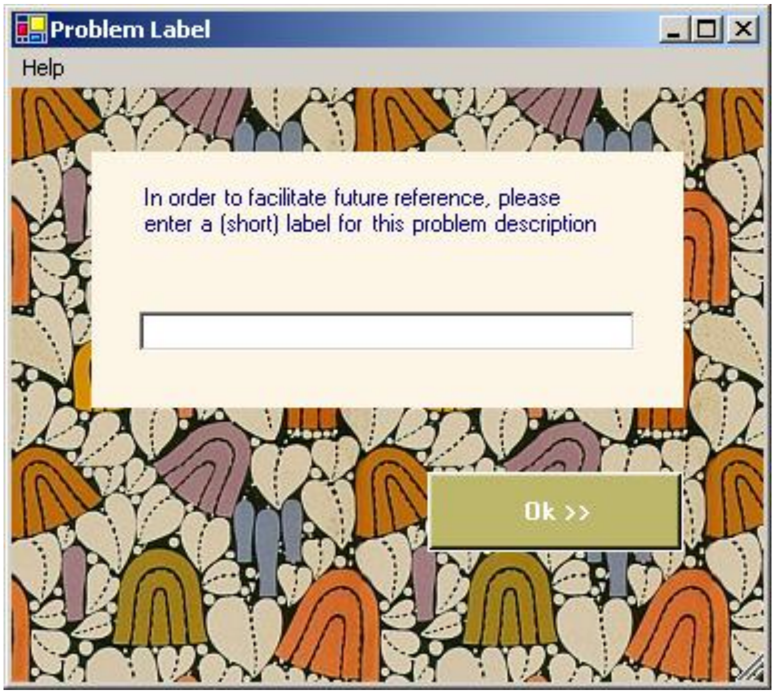
Finally, a Problem Reviewal form allows a final check of all inputs, and to select an html output file location.

3.1 Saving for later


Run Now
Proceed to sample size calculation >>
Run Later
Register Problem Description >>

The bottom right buttons of the Problem Reviewal form, illustrated at the right, allows the user to launch the sample size calculation (Run Now) or to save the problem description (to an internal file called a script), which you will launch when you are ready, e.g., when you have finished entering a series of **SSCOR** scripts, each with different prior distributions or sample size criterion. If you are just running a single sample size calculation, you will typically want to click on "Run Now". The program will then begin to run to find the optimal sample size for your inputs.

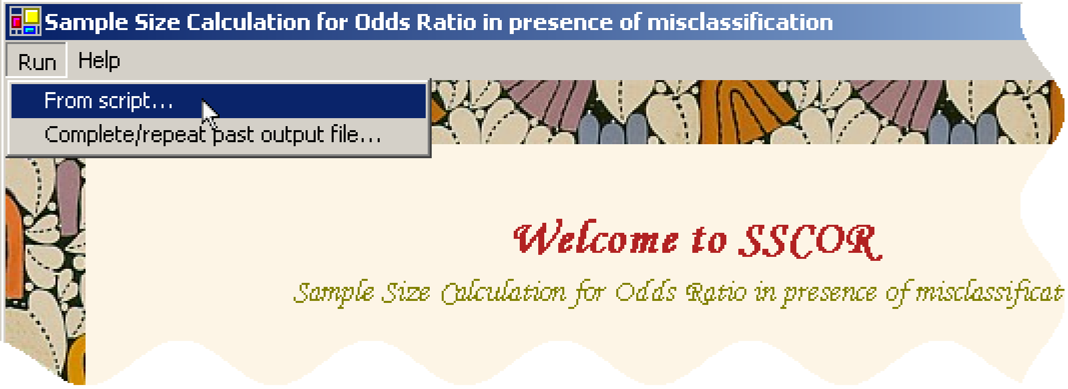
In order to make a script easily recognizable, you will be asked to enter a label for the problem entered.



In order to facilitate future reference, please enter a (short) label for this problem description

3.1.1 Running scripts

If you want to run a series of **SSCOR** scripts, open **SSCOR** and select the **Run/From Script** item from the top menu of the first form


Sample Size Calculation for Odds Ratio in presence of misclassification
Run
Help
From script…
Complete/repeat past output file…

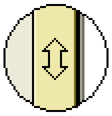
Welcome to SSCOR

Sample Size Calculation for Odds Ratio in presence of misclassificat

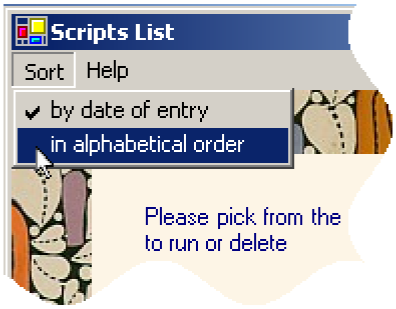
You can select only one script to run or a subset of the scripts from the list, or all of them. They can be deleted when the program completes or not, depending on whether or not you check the **Delete script(s) after completion** tick box.

If you cannot exactly remember of the problem saved under any of the script label in the list, select that label and click the bottom right **View** button.

Script labels are displayed in order of date of entry.



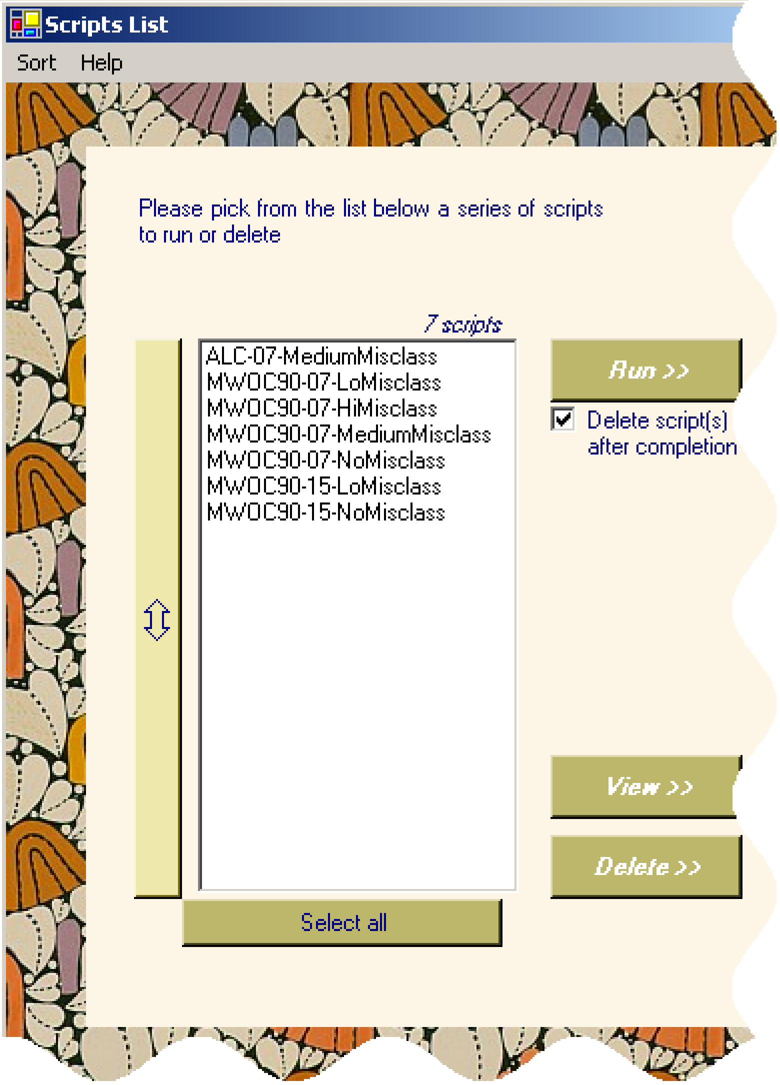
The double-sided arrow button will sort the labels in ascending/descending order, alternatively.



Please pick from the to run or delete

You can also sort entries in alphabetical order by changing the sort key from the top-left **Sort** menu item.

Selected scripts will be submitted in the order in which they appear, from top to bottom, when **Run>>** button is clicked.



Please pick from the list below a series of scripts to run or delete

7 scripts

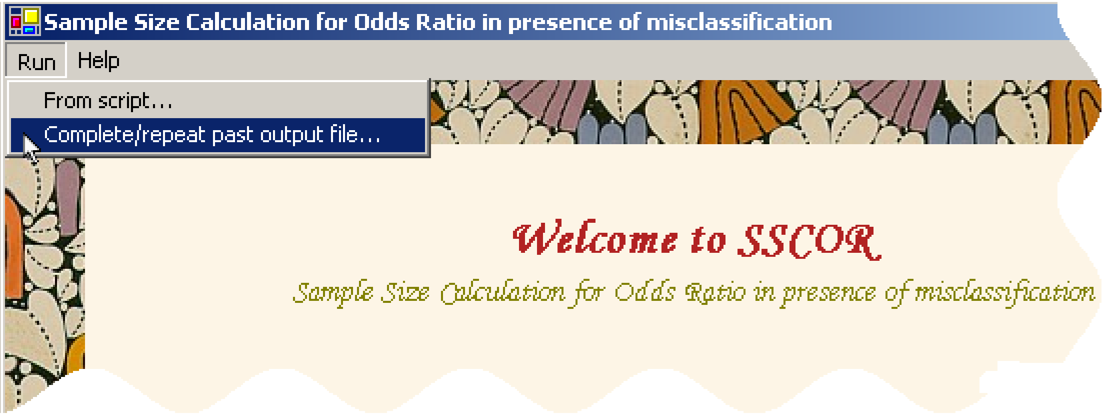
ALC-07-MediumMisclass
MWOC90-07-LoMisclass
MWOC90-07-HiMisclass
MWOC90-07-MediumMisclass
MWOC90-07-NoMisclass
MWOC90-15-LoMisclass
MWOC90-15-NoMisclass

☑ Delete script(s) after completion

3.2 Resuming previously stopped sample size calculations

SSCOR can take a long time to run, depending on the various inputs. Sometimes you may wish to stop calculation temporarily, and continue later.

A **SSCOR** calculation can be resumed by reloading the interrupted **SSCOR** html output file from the **Run/Complete-Repeat past output file** menu from **SSCOR** initial form.

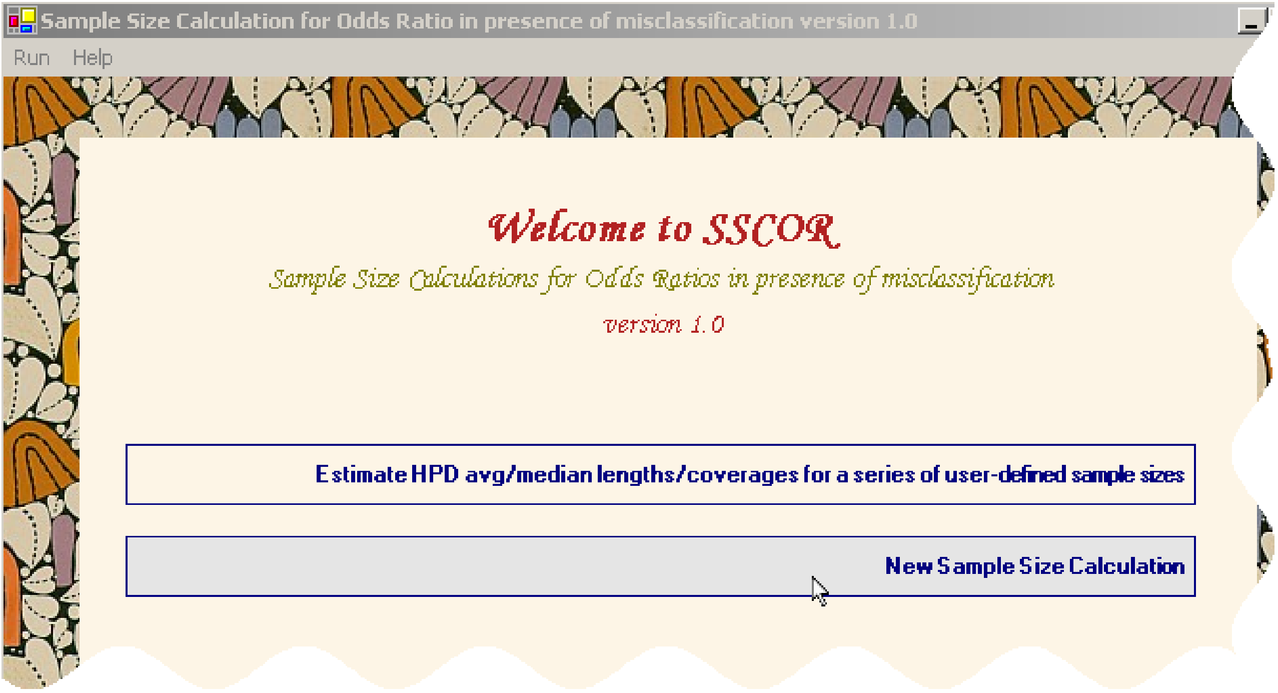

Sample Size Calculation for Odds Ratio in presence of misclassification
Run
Help
From script...
Complete/repeat past output file...
Welcome to SSCOR
Sample Size Calculation for Odds Ratio in presence of misclassification

4. An example of running SSCOR

We will illustrate the use of **SSCOR** to estimate sample size requirements with a problem from Table 2 of our paper *Bayesian sample size determination for case-control studies when exposure may be misclassified,* cited in the introduction.

Consider a scenario where the true OR is around 0.7 and a HPD interval of fixed length 0.4 is desired. Consider further a moderate misclassification rate, with a narrow prior around this rate. We will use the ACC criterion to compute the minimal sample size such that average HPD coverage is 95%.

Click the *New Sample Size Calculation* item menu from the initial form.


Sample Size Calculation for Odds Ratio in presence of misclassification version 1.0
Run
Help
Welcome to SSCOR
Sample Size Calculations for Odds Ratios in presence of misclassification
version 1.0
Estimate HPD avg/median lengths/coverages for a series of user-defined sample sizes
New Sample Size Calculation

`

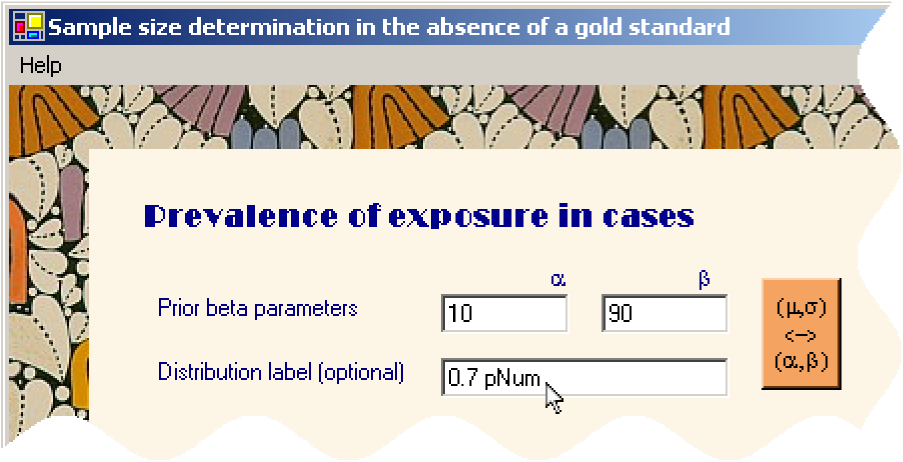
We assume a Beta(10, 90) prior for the exposure rate within the case group, and a Beta (13.7, 86.3) prior for the exposure rate within the control group. Note that these prior densities have means of 0.1 and 0.137, respectively, and when random paired values are chosen from each density, ORs center at about 0.7.

Enter 10 and 90 in the α and β text boxes, respectively.

Enter a label in the Distribution label text box: we have entered a label (0.7 pNum) for the prior distribution used.

Doing so will make this uniform prior only one click away the next time you run this program.

Click *Next>>* button when done.


Sample size determination in the absence of a gold standard

Prevalence of exposure in cases

Prior beta parameters α 10 β 90

Distribution label (optional) 0.7 pNum

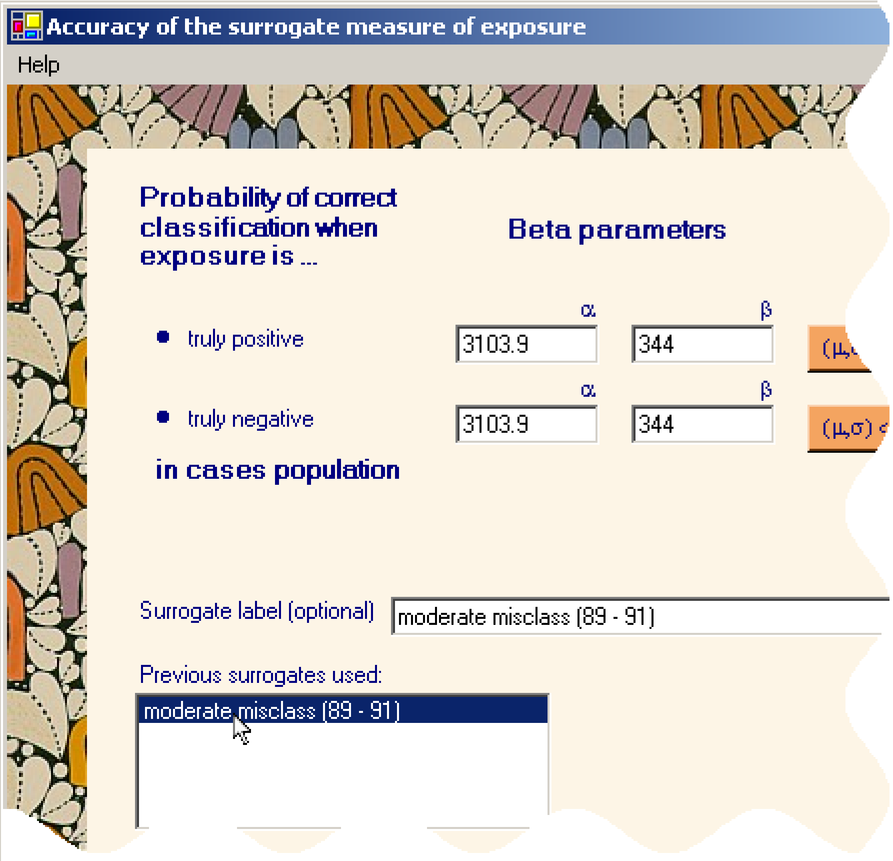
Enter 13.7 and 86.3 for α and β in next form, labeled Prevalence of exposure in controls. Enter a distribution label of you plan to reuse that prior distribution in future **SSCOR** sample size calculations.

Consider the scenario where we know relatively well the rate of misclassification, which is moderate, that is, where was assume the misclassification to be in the 9-11 % range.

In the form illustrated here, we take advantage of the fact that this prior distribution has been used in the past and labeled it *moderate misclass (89-91).* Clicking that label in *Previous surrogates used* list box in the left lower part of the form automatically fills the boxes with the appropriate parameter values.

In your case, for your first use of **SSCOR**, you will need to fill in the boxes with the appropriate values, as illustrated here.

Click *Next>>* button.



Accuracy of the surrogate measure of exposure

Probability of correct classification when exposure is ...

Beta parameters

	α	β
● truly positive	3103.9	344
● truly negative	3103.9	344

in cases population

Surrogate label (optional): moderate misclass (89 - 91)

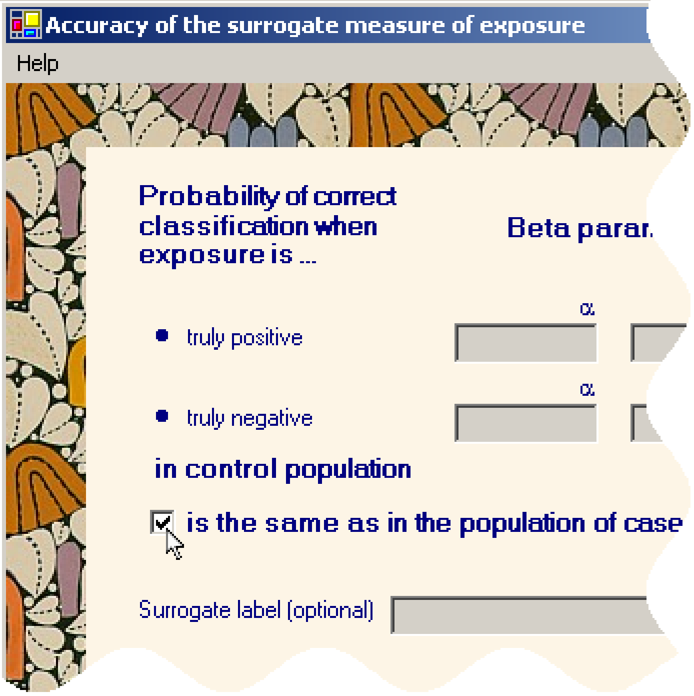
Previous surrogates used:

moderate misclass (89 - 91)

Enter the misclassification rate within the control population prior distribution through next form. Prior distribution for misclassification in controls population could very well be different from that of cases population misclassification rate.

In this example, we will assume the misclassification rates in control and case populations are **exactly the same.** Check the appropriate tick box to indicate that rates are identical in both populations. If you want to use different values, they are filled in as in the examples above.

Click *Next>>* button.



Accuracy of the surrogate measure of exposure

Probability of correct classification when exposure is ...

Beta parar

	α
● truly positive	
● truly negative	

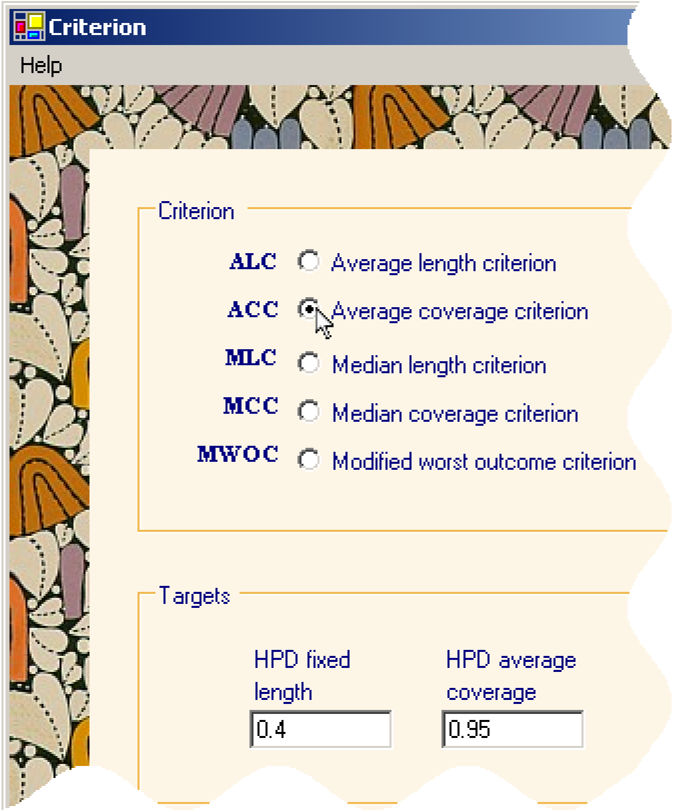
in control population

☑ is the same as in the population of case

Surrogate label (optional)

In next form, select the Average coverage criterion and enter the fixed HPD length in the relevant text box, as illustrated.

Click *Next>>* button.



Criterion

- ALC ○ Average length criterion
- ACC ● Average coverage criterion
- MLC ○ Median length criterion
- MCC ○ Median coverage criterion
- MWOC ○ Modified worst outcome criterion

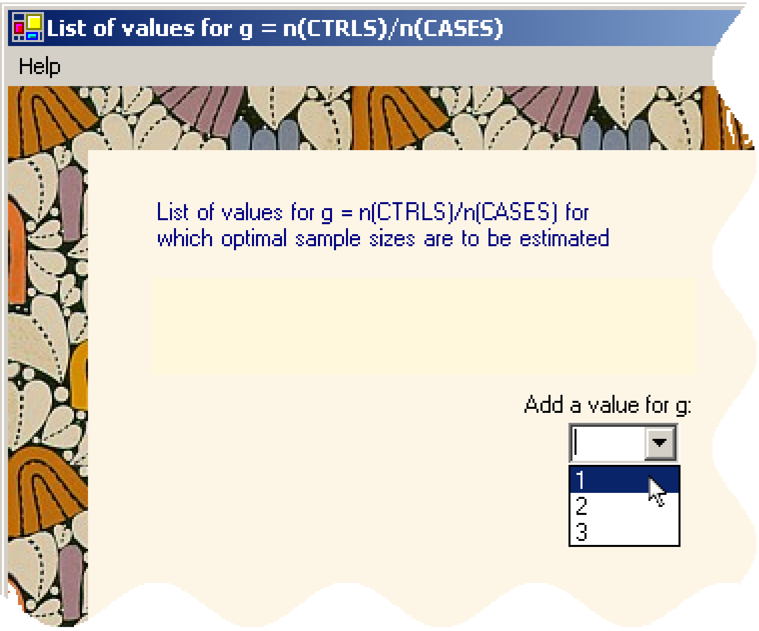
Targets

HPD fixed length	HPD average coverage
0.4	0.95

The next form allows you to specify the control-to- case ratio, that is, the ratio *g* = number of controls sampled divided by the number of cases sampled.

You can pick a value from the suggested list or enter your own value for *g*. More than one value can be selected for *g*.

Select *g* = 1 and click Next>> button when done.



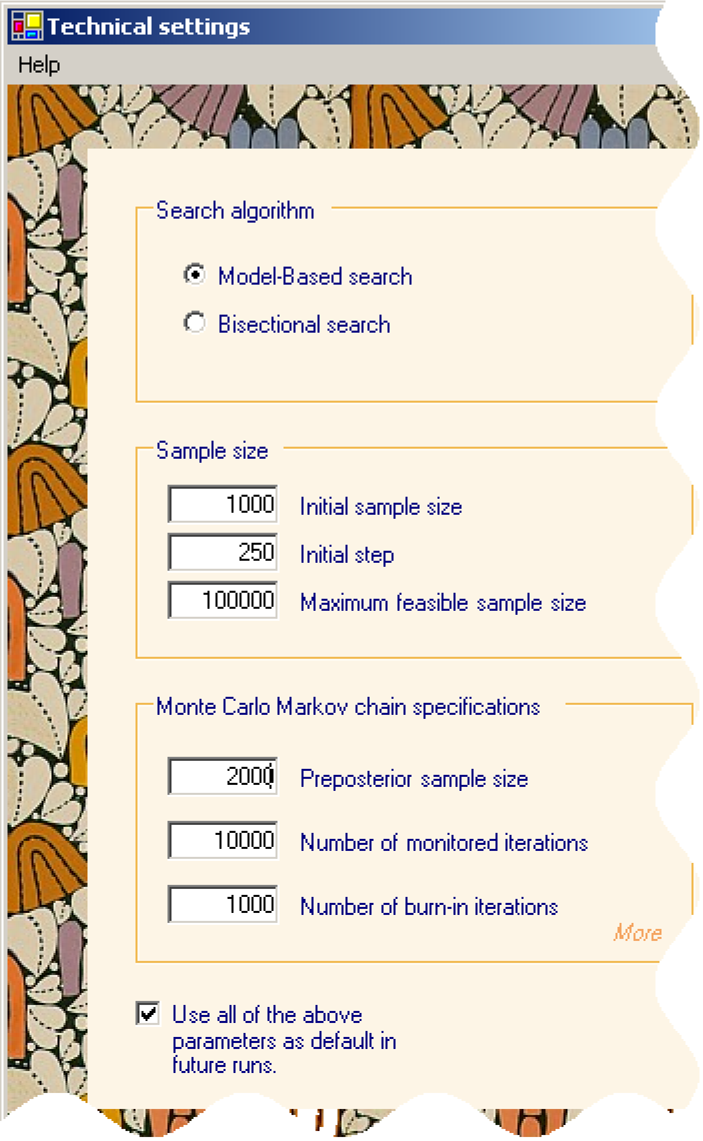
List of values for g = n(CTRLS)/n(CASES) for which optimal sample sizes are to be estimated

Add a value for g:

1
2
3

In this example, we will accept the default settings except for the preposterior sample size, which we will double to 2000 to derive a more accurate outcome estimate at the expense of slightly longer running time. You can select these values to be the default in future **SSCOR** utilizations by checking the appropriate tick box at the bottom of the form.

Click *Next>>* button when done.



Technical settings

Search algorithm

● Model-Based search

○ Bisectional search

Sample size

1000	Initial sample size
250	Initial step
100000	Maximum feasible sample size

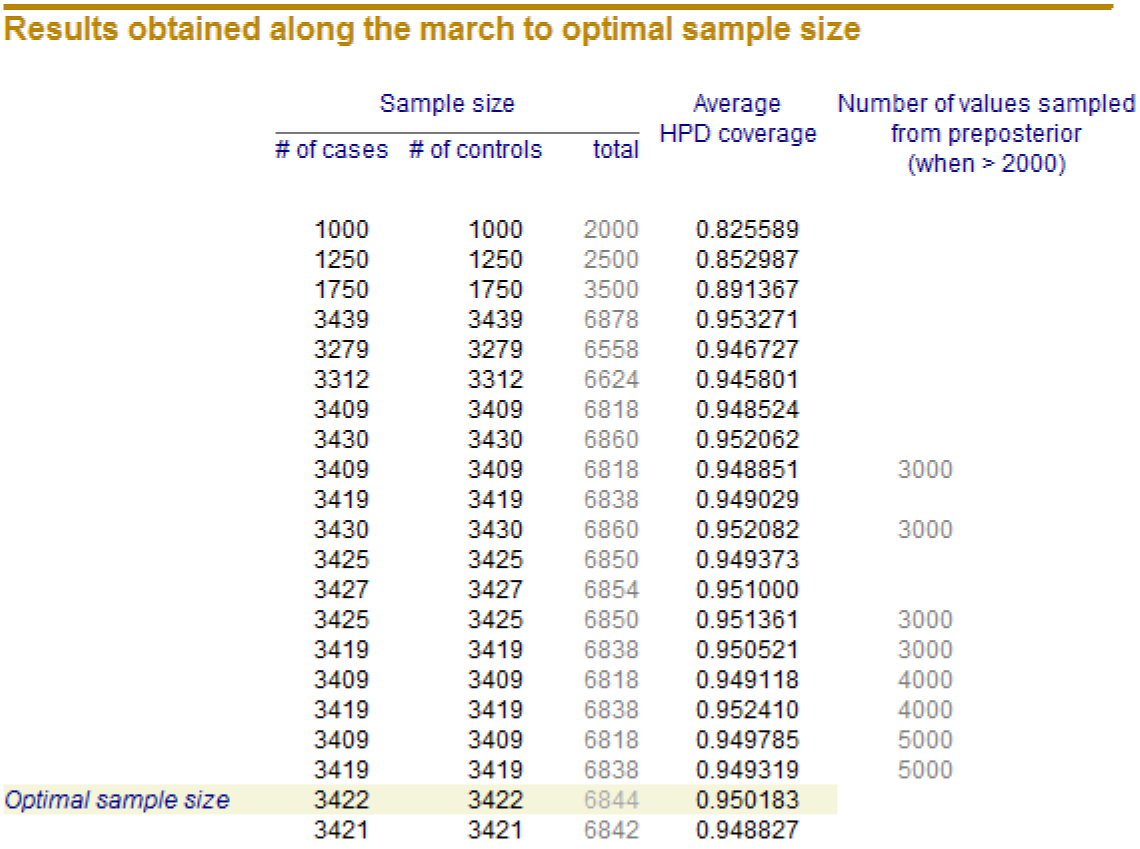
Monte Carlo Markov chain specifications

2000	Preposterior sample size
10000	Number of monitored iterations
1000	Number of burn-in iterations

☑ Use all of the above parameters as default in future runs.

Last but not least, the Problem Reviewal form allows you to view at a glance and revisit each parameter value entered in your sample size problem description. As already discussed in Sections 3 and 3.1, you will select the html output file location through this final form and save the problem description to as script for future run or launch the calculation right away.

The final html output file will have a section presenting the intermediate sample sizes along with the outcome estimated for each visited sample size (next page).



Results obtained along the march to optimal sample size

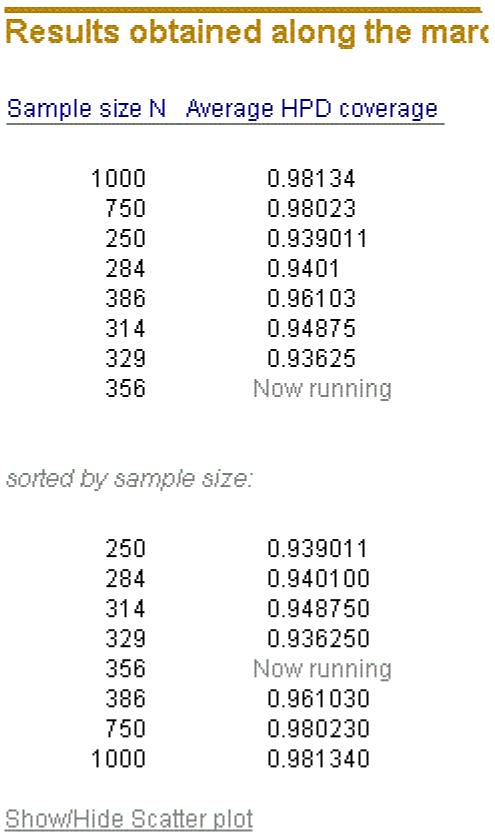
	Sample size			Average HPD coverage	Number of values sampled from preposterior (when > 2000)
	# of cases	# of controls	total		
	1000	1000	2000	0.825589	
	1250	1250	2500	0.852987	
	1750	1750	3500	0.891367	
	3439	3439	6878	0.953271	
	3279	3279	6558	0.946727	
	3312	3312	6624	0.945801	
	3409	3409	6818	0.948524	
	3430	3430	6860	0.952062	
	3409	3409	6818	0.948851	3000
	3419	3419	6838	0.949029	
	3430	3430	6860	0.952082	3000
	3425	3425	6850	0.949373	
	3427	3427	6854	0.951000	
	3425	3425	6850	0.951361	3000
	3419	3419	6838	0.950521	3000
	3409	3409	6818	0.949118	4000
	3419	3419	6838	0.952410	4000
	3409	3409	6818	0.949785	5000
	3419	3419	6838	0.949319	5000
Optimal sample size	3422	3422	6844	0.950183	
	3421	3421	6842	0.948827	

In this example, average HPD coverage was first estimated for N = 1000, our initial sample size. Since that sample size was too small, it was increased by 250, our initial step value. Since N = 1250 was still insufficient, the sample size was increased by twice the initial step, that is, by an amount of 500, to get N = 1750 in the third average HPD coverage estimation.

Since we have selected a model based sample size, the remaining steps are based on results from the (maximum 10) previous steps, where the best-fitting curve is used to estimate optimal sample size. The sample size search was stopped after average the HPD coverage for N=3421 was estimated, since it was shown to be an insufficient sample size while N=3422 was previously shown to be sufficient. Of course, our program uses WinBUGS, which has random Monte Carlo error, so that re-running the program may result in slightly different sample sizes, but it should be close.

4.1 Iterative html output file updates

Whether you are running **SSCOR** for a sample size calculation or for outcome estimation for a series of predetermined sample sizes, you may be interested in having a look at intermediate results while **SSCOR** is running. The main html output file is updated after the outcome of interest is estimated for each sample size, and viewing it in your favorite browser is possible at any point in time.



Results obtained along the marc

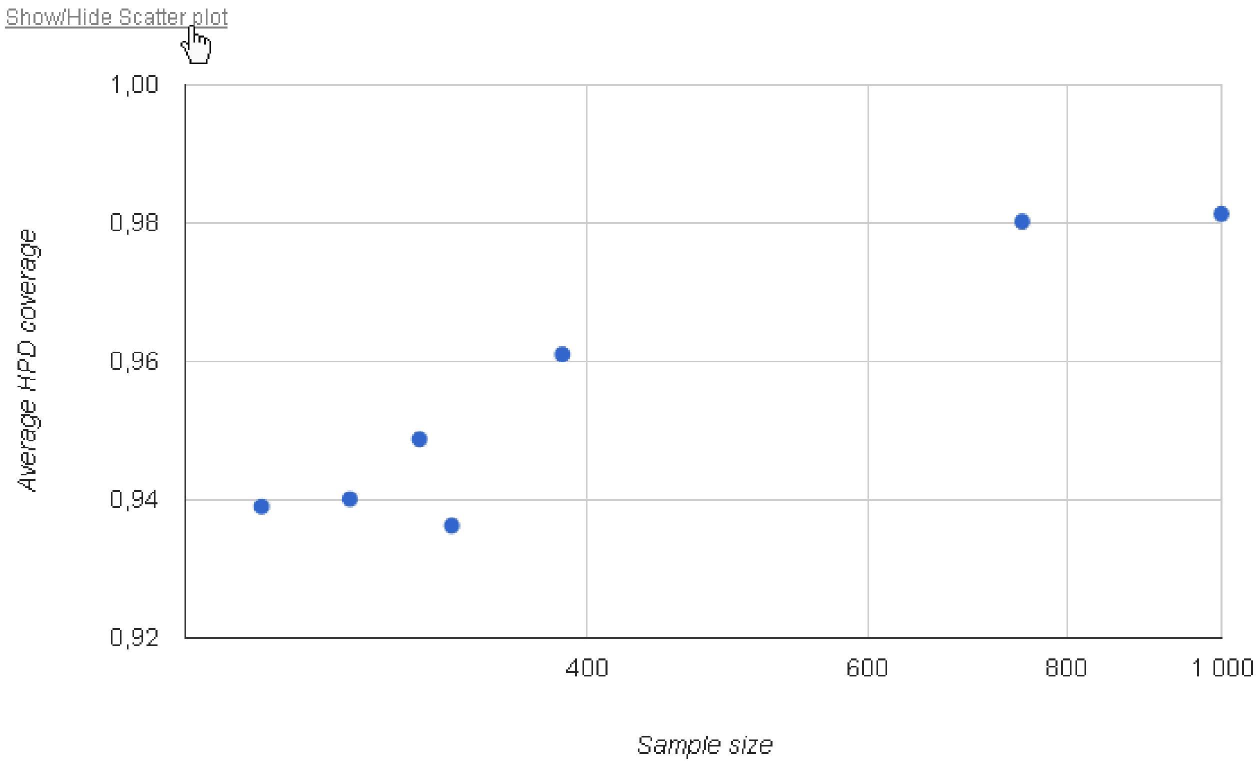
Sample size N	Average HPD coverage
1000	0.98134
750	0.98023
250	0.939011
284	0.9401
386	0.96103
314	0.94875
329	0.93625
356	Now running

sorted by sample size:

250	0.939011
284	0.940100
314	0.948750
329	0.936250
356	Now running
386	0.961030
750	0.980230
1000	0.981340

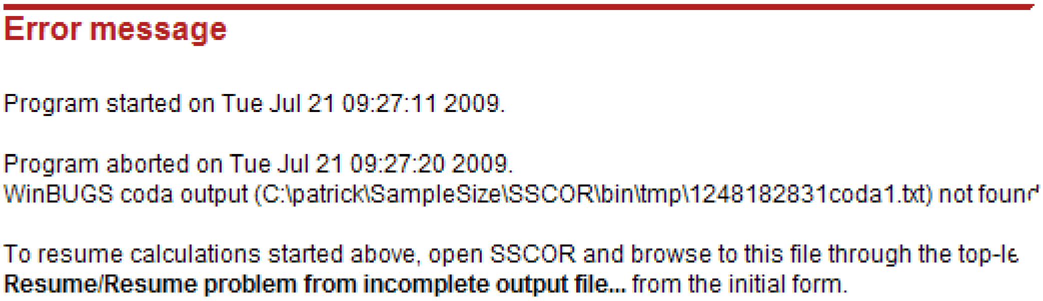
The output section labeled **Results obtained along the march to optimal sample size** is divided into two subsections: the upper portion lists the series of sample sizes along with their corresponding estimated outcome, in the order in which they were run. This is followed by a subsection where the same results are listed in ascending order of sample size, which may be easier to follow. Note that if the outcome had to be re-estimated for a given sample size (along the search for optimal sample size), that sample size will appear two or more times in the upper section, once per re-estimation, while only the final estimate (which summarizes information from each intermediate outcome estimate) will appear in the lower (sorted) section.

Finally, a link labeled *Show/Hide Scatter plot* opens a scatter plot (see example below) which helps visualize the relationship between outcome and sample size; it may even give enough information to the user with regards to the final sample size, even though the program has not formally reached convergence, and the user may decide to stop the program before it does converge. Note that when **SSCOR** finishes, a .pdf document presents the same scatter plot with a little bit more information.



5. Error messages

If **SSCOR** fails during a sample size calculation, an error message will appear at the bottom of the html output file.



Error message

Program started on Tue Jul 21 09:27:11 2009.

Program aborted on Tue Jul 21 09:27:20 2009.
WinBUGS coda output (C:\patrick\SampleSize\SSCOR\bin\tmp\1248182831coda1.txt) not found

To resume calculations started above, open SSCOR and browse to this file through the top-le
Resume/Resume problem from incomplete output file... from the initial form.

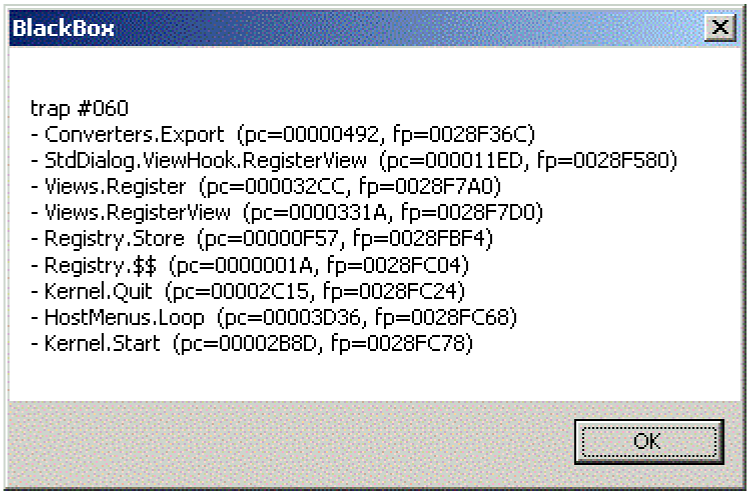
However, if it fails at an earlier stage, that is, in code prior to actual sample size calculations, the error message will only appear in a text file called **SSCOR.txt**, found in a **log/** subdirectory of the main directory where **SSCOR** was installed (C:\Users\user name\Documents\Bayesian Software\SSCOR or C:\Documents and Settings\user name\My Documents\Bayesian Software\SSCOR , by default, depending on your platform).

If an html file was created, you may want to resume the sample size calculations (see section 3.2). If you do not succeed in resuming, do not hesitate to communicate with us (Lawrence.Joseph@McGill.ca).

6. Avoiding trap errors from permission settings 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 **SSCOR** to run smoothly.



```
trap #060
- Converters.Export  (pc=00000492, fp=0028F36C)
- StdDialog.ViewHook.RegisterView  (pc=000011ED, fp=0028F580)
- Views.Register  (pc=000032CC, fp=0028F7A0)
- Views.RegisterView  (pc=0000331A, fp=0028F7D0)
- Registry.Store  (pc=00000F57, fp=0028FBF4)
- Registry.$$  (pc=0000001A, fp=0028FC04)
- Kernel.Quit  (pc=00002C15, fp=0028FC24)
- HostMenus.Loop  (pc=00003D36, fp=0028FC68)
- Kernel.Start  (pc=00002B8D, fp=0028FC78)
```

7..Change log

Versions 1.2 and 1.2.1 (February 2014)
We have improved the stopping criterion to avoid very long runs of the program (see section 2.2). We have also made more efficient use of information from previous runs when repeating an outcome estimate for a given simple size. Finally, scatter plot of outcome vs sample size is now embedded in the main html output file.

Versions 1.3 -- 1.3.2 (April 2014)
Minor improvement to the model-based search algorithm.

Version 1.4 (April 2014)
Minor update.

Versions 1.5 -- 1.5.2 (May 2014)
Added automated sequences for user-defined sample sizes for which outcome is to be estimated.

Version 1.6 (January 2015)
Minor update.

Versions 1.6.1 -- 1.6.5 (April 2015)
Minor bug fix update: a potential installation problem was fixed.

Version 1.7 (January 2016)
Minor update.

Version 1.8 (September 2017)
SSCOR now works on Windows 8 & 10. Windows 7 users do not need to reinstall or upgrade.

Version 1.8.1 (February 2019)
Depending on user's R settings, some temporary R output files were not saved along the expected lines, causing **SSCOR** to crash in previous version. Problem solved.

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

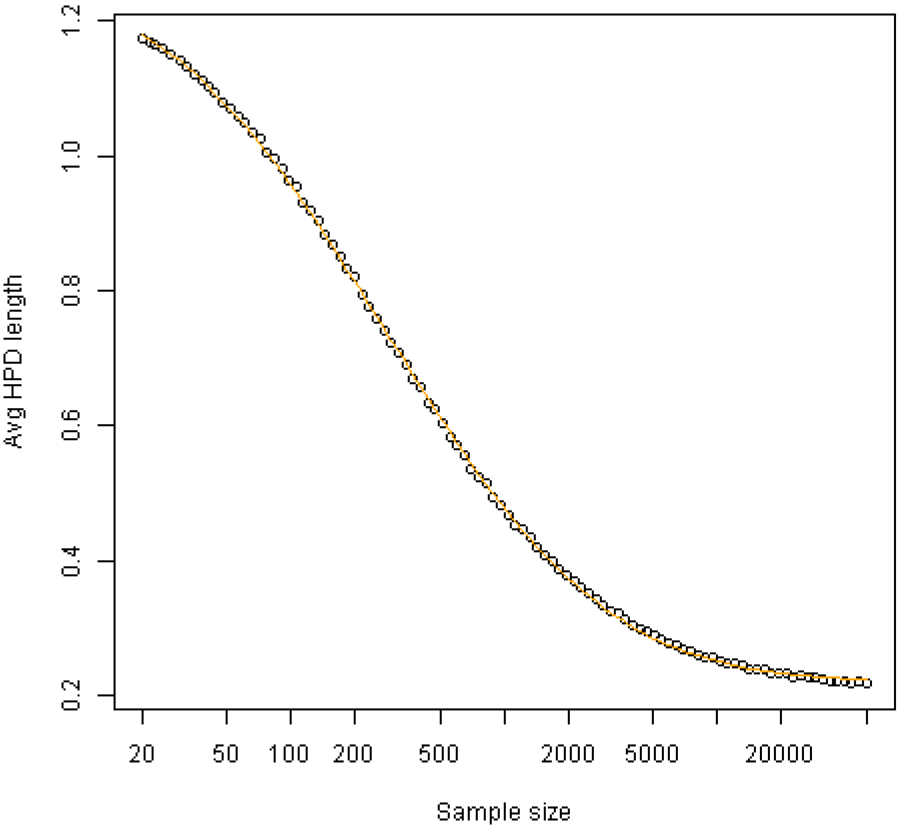
Appendix A
Model based strategy in sample size calculations

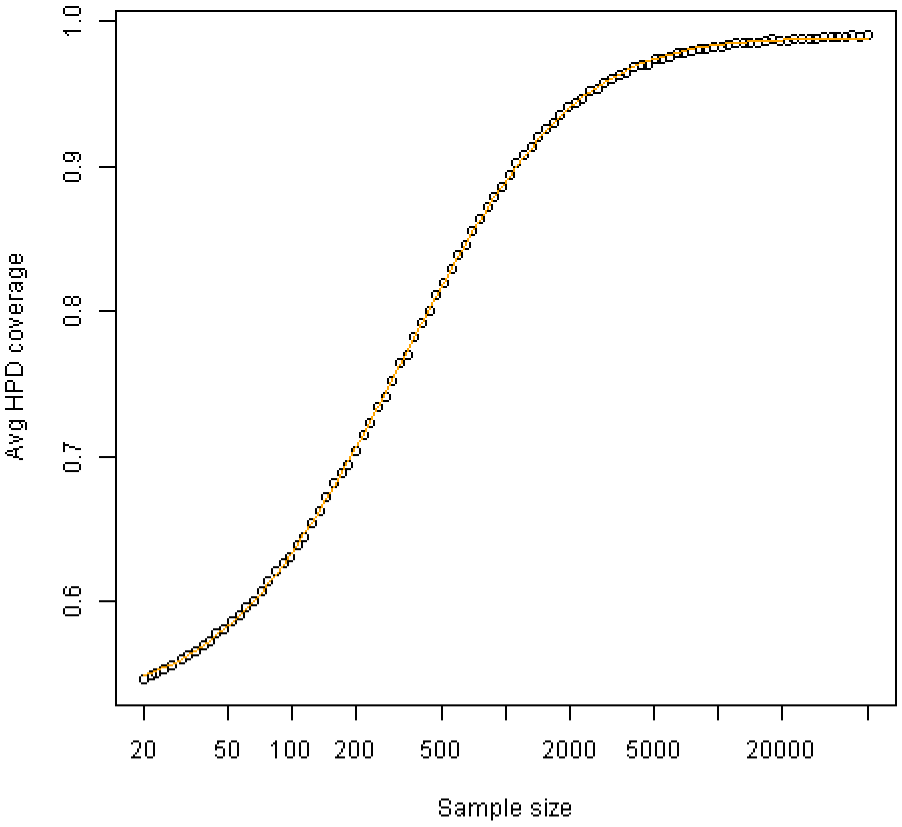
Sample size calculations run in **SSCOR** can follow a bisectional or model based path towards the optimal sample size. The equation under the model based search is

$$\text{outcome} = a + b\,\Phi(\,(\log(N) - \mu) / \sigma) \tag{1}$$

where Φ is the cumulative normal density. In practice, we have found that the equation fits the data very well in most situations.

The opposite figure and the figure below present the average HPD length and average HPD coverage vs different equally-spaced sample sizes (on the log scale) obtained from simulation (where preposterior sample size was chosen large enough to have MCMC error significantly reduced) for two different problems, along with the fitted curve, showing an excellent fit. This is not a specially selected example, but is quite typical of the fits our model provides.





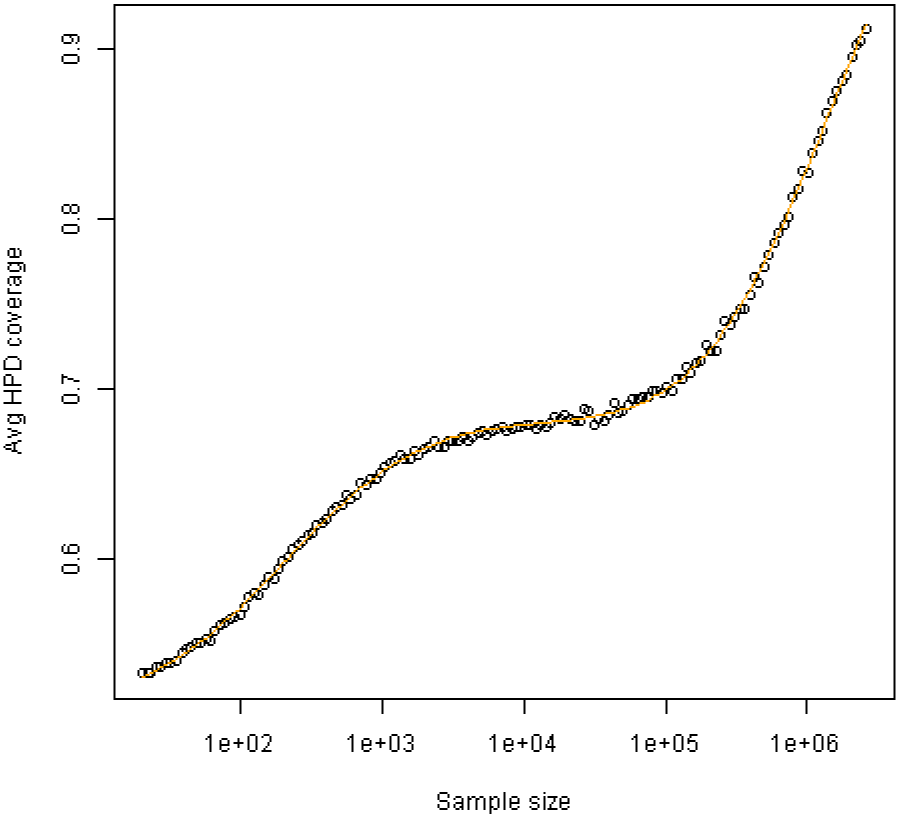
When too few points were visited in the march towards optimal sample size to estimate the above equation parameters, a monotonic third-degree polynomial or a second-degree polynomial (still in terms of log(N)) will be estimated.

In some cases, we have also found the equation

$$\text{outcome} = a + p\, b\, \Phi(\,(\log(N) - \mu_1) / \sigma_1) + (1\text{-}p)\, b\, \Phi(\,(\log(N) - \mu_2) / \sigma_2) \qquad (2)$$

to better fit the data than its one-term version (1). This equation requires more points at hand to be estimated and is thus used only when the outcome was estimated for 20 sample sizes or more. When **SSCOR** is used for sample size calculation, however, only the last 10 points along the walk towards optimal sample size are retained and hence the previous equation is never used in that context; it will be used only in the context of outcome estimation for (a large number of) pre-specified sample sizes.

The opposite figure shows the fit of the binormal equation above to average HPD coverage obtained for a large number of sample sizes in presence of moderate misclassification.


Avg HPD coverage
1e+02
1e+03
1e+04
1e+05
1e+06
Sample size