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, Belisle P

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. 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\username\Documents\Bayesian Software\SSCOR or C:\Documents and Settings\username\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

<table>
<thead>
<tr>
<th></th>
<th>Cases ($D_1$)</th>
<th>Controls ($D_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>$E_0$</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>$n_1$</td>
<td>$n_0$</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Cases ($D_1$)</th>
<th>Controls ($D_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$</td>
<td>$\theta_1$</td>
<td>$\theta_0$</td>
</tr>
<tr>
<td>$E_0$</td>
<td>$1-\theta_1$</td>
<td>$1-\theta_0$</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

where $\theta_0$ and $\theta_1$ 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

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

We thus observe the cell counts

<table>
<thead>
<tr>
<th></th>
<th>Cases ($D_1$)</th>
<th>Controls ($D_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E^* = 1$</td>
<td>$X_1$</td>
<td>$X_0$</td>
</tr>
<tr>
<td>$n_1$</td>
<td>$n_0$</td>
<td>$N$</td>
</tr>
<tr>
<td>$E^* = 0$</td>
<td>$n_1-X_1$</td>
<td>$n_0-X_0$</td>
</tr>
</tbody>
</table>

with disease conditional probabilities of apparent (that is, measured by an imperfect surrogate) exposure
<table>
<thead>
<tr>
<th>Cases ($D_1$)</th>
<th>Controls ($D_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1^*$</td>
<td>$\theta_0^*$</td>
</tr>
<tr>
<td>$1-\theta_1^*$</td>
<td>$1-\theta_0^*$</td>
</tr>
</tbody>
</table>

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 $\theta_i$ 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.

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 $(\alpha, \beta)$, such that prior mean and variance are $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha \beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively.

The orange button with text $(\mu, \sigma) \leftrightarrow (\alpha, \beta)$ allows you to specify your prior distributions in terms of prior moments $(\mu, \sigma)$ rather than in terms of beta parameters $(\alpha, \beta)$. If you choose to enter your prior information using $(\mu, \sigma)$, the corresponding $(\alpha, \beta)$ 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.

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.
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.
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.

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

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.
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.

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.

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.
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.

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%.
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 $\alpha$ and $\beta$ 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.

Enter 13.7 and 86.3 for $\alpha$ and $\beta$ 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 we 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.

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

Click Next>> button.

The next form allows you to specify the control-to-case ratio, that is, the ratio \( g = \frac{\text{number of controls sampled}}{\text{number of cases sampled}} \).

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.
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.

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

<table>
<thead>
<tr>
<th>Sample size # of cases</th>
<th># of controls</th>
<th>total</th>
<th>Average HPD coverage</th>
<th>Number of values sampled from preposterior (when &gt; 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>1000</td>
<td>2000</td>
<td>0.825589</td>
<td></td>
</tr>
<tr>
<td>1250</td>
<td>1250</td>
<td>2500</td>
<td>0.852987</td>
<td></td>
</tr>
<tr>
<td>1750</td>
<td>1750</td>
<td>3500</td>
<td>0.891367</td>
<td></td>
</tr>
<tr>
<td>3439</td>
<td>3439</td>
<td>6878</td>
<td>0.953271</td>
<td></td>
</tr>
<tr>
<td>3279</td>
<td>3279</td>
<td>6558</td>
<td>0.946727</td>
<td></td>
</tr>
<tr>
<td>3312</td>
<td>3312</td>
<td>6624</td>
<td>0.945801</td>
<td></td>
</tr>
<tr>
<td>3409</td>
<td>3409</td>
<td>6818</td>
<td>0.948524</td>
<td>3000</td>
</tr>
<tr>
<td>3430</td>
<td>3430</td>
<td>6860</td>
<td>0.952062</td>
<td></td>
</tr>
<tr>
<td>3409</td>
<td>3409</td>
<td>6818</td>
<td>0.943851</td>
<td></td>
</tr>
<tr>
<td>3419</td>
<td>3419</td>
<td>6838</td>
<td>0.949029</td>
<td></td>
</tr>
<tr>
<td>3430</td>
<td>3430</td>
<td>6860</td>
<td>0.952082</td>
<td>3000</td>
</tr>
<tr>
<td>3425</td>
<td>3425</td>
<td>6850</td>
<td>0.949373</td>
<td></td>
</tr>
<tr>
<td>3427</td>
<td>3427</td>
<td>6854</td>
<td>0.951000</td>
<td></td>
</tr>
<tr>
<td>3425</td>
<td>3425</td>
<td>6850</td>
<td>0.951361</td>
<td>3000</td>
</tr>
<tr>
<td>3419</td>
<td>3419</td>
<td>6838</td>
<td>0.950521</td>
<td>3000</td>
</tr>
<tr>
<td>3409</td>
<td>3409</td>
<td>6818</td>
<td>0.949118</td>
<td>4000</td>
</tr>
<tr>
<td>3419</td>
<td>3419</td>
<td>6838</td>
<td>0.952410</td>
<td>4000</td>
</tr>
<tr>
<td>3409</td>
<td>3409</td>
<td>6818</td>
<td>0.949785</td>
<td>5000</td>
</tr>
<tr>
<td>3419</td>
<td>3419</td>
<td>6838</td>
<td>0.949319</td>
<td>5000</td>
</tr>
</tbody>
</table>

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 march to optimal sample size

Sample size N  Average HPD coverage

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Average HPD coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.98134</td>
</tr>
<tr>
<td>750</td>
<td>0.98023</td>
</tr>
<tr>
<td>250</td>
<td>0.939011</td>
</tr>
<tr>
<td>284</td>
<td>0.9401</td>
</tr>
<tr>
<td>386</td>
<td>0.96103</td>
</tr>
<tr>
<td>314</td>
<td>0.94875</td>
</tr>
<tr>
<td>329</td>
<td>0.93625</td>
</tr>
<tr>
<td>356</td>
<td>Now running</td>
</tr>
</tbody>
</table>

sorted by sample size:

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Average HPD coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0.939011</td>
</tr>
<tr>
<td>284</td>
<td>0.940100</td>
</tr>
<tr>
<td>314</td>
<td>0.948750</td>
</tr>
<tr>
<td>329</td>
<td>0.936250</td>
</tr>
<tr>
<td>356</td>
<td>Now running</td>
</tr>
<tr>
<td>386</td>
<td>0.961030</td>
</tr>
<tr>
<td>750</td>
<td>0.980230</td>
</tr>
<tr>
<td>1000</td>
<td>0.981340</td>
</tr>
</tbody>
</table>

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\1248182831\coda1.txt) not found

To resume calculations started above, open SSCOR and browse to this file through the top-level 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.

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.

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

\[ \text{outcome} = a + b \Phi(\frac{\log(N) - \mu}{\sigma}) \]  

(1)

where \( \Phi \) 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

\[
\text{outcome} = a + p \cdot b \cdot \Phi\left(\frac{\log(N) - \mu_1}{\sigma_1}\right) + (1-p) \cdot b \cdot \Phi\left(\frac{\log(N) - \mu_2}{\sigma_2}\right)
\]

(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.