

INTRODUCING THE **ESPRESSO** POWER CALCULATOR

PREAMBLE

Assessment error in outcome or explanatory variables can substantially reduce the power of an association analysis. **ESPRESSO** (Estimating Sample-size and Power in R by Exploring Simulated Study Outcomes) is a new software program (implemented in the statistical environment “R”) that supports simulation-based power calculation for stand-alone case-control studies and for case-control analyses nested in cohort studies, that take account of realistic assessment error. The original development of this program was funded by UK Biobank and an early version of the program was used to carry out the power calculations for that project.

Simulation based power calculation involves repeatedly simulating a data set with a number of key characteristics and seeing in what proportion of the simulations the effect of interest can be detected using whatever approach to analysis is to be used in practice. **ESPRESSO.CC** focuses on models with a binary outcome (case=1, control=0) – *i.e.* case-control studies. The version of **ESPRESSO** that is currently implemented addresses the lowest power setting that is widely encountered, that is, when the exposure variables are also binary: a life-style risk factor (present=1, absent=0) and a genetic risk factor (present=1, absent=0). Analyses such as these are often power limiting and therefore sample size determining for a stand-alone or nested case-control study. **ESPRESSO-CC** extends an earlier version of **ESPRESSO** by including a switch (input parameter #6) that allows the model to include, or not include, an interaction between the genetic and environmental determinant. It also includes a second switch (input parameter #7) that allows the binary genetic model to be replaced by an additive genetic model (*i.e.* the covariate is “the number of minor alleles”). At present **ESPRESSO** does not deal with model misspecification (*e.g.* if the true genetic model is binary but it is analysed as additive, or *vice-versa*) – this will follow later. All modelling (both the simulation and analysis steps – see IJE paper and supplementary methods) is based on unconditional logistic regression.

In order to run this simulation-based power calculator you have to specify a series of input parameters. In the interactive version of the program that is run directly from the **P³G** website, these parameters are specified by typing numbers into appropriate boxes. The program is then run by clicking the “Run calculation” button at the bottom of the screen and the required answers come up in an output window. More detailed output is also produced and can be downloaded as .csv (semi-colon delimited) Excel files. In the batch mode version, the input parameters are specified in a semi-colon (;) delimited .csv file – this enables many different simulation scenarios to be specified and run together, all in one session. The input and output parameters are described in detail below.

INPUT PARAMETERS

#1. Scenario ID

A integer code for each simulation scenario that enables them to be identified in the output. There is no requirement that the IDs be sequential.

#2. Random number seed

Simulation based power calculation involves generating random numbers. This is done by using a “pseudorandom number generator”. This generates a sequence of numbers that behaves as though it is random, even though it is actually deterministic. Because it is deterministic, once a specific starting “seed”

has been specified, it is then possible to generate precisely the same sequence again by using the same seed. If you specify a seed and keep a record of that seed, it means that you can then repeat precisely the same analysis again if, for example, you want to look at the results in more detail, or to check the results if you didn't record the output correctly. In ESPRESSO you need to specify a seed that is a whole number between 1 and 9999999.

#3. The number of simulations

The more simulations that are run, the better reflection the set of simulations will provide of the particular scenario that has been specified. At present, the version of **ESPRESSO** on the **P³G** server is limited to running 500 simulations because of limitations on time and computer load. When you first start exploring a study design you may well wish to explore a broad range of scenarios relatively quickly and so you may choose to run 100 or even 50 simulations. But when you come to produce definitive results for a power table, we suggest that you run at least 500 simulations. If you download the ESPRESSO program and run it on your own server the number of simulations is unlimited except by memory (you may need to reset the amount of memory available to R). In the implemented version of the program, the default number of simulations is set to 100.

#4. The number of cases, and #5. the number of controls

Here, there are two different, but important, issues: (1) **The ratio of the number of cases to controls.** If you specify 250 cases and 750 controls, part of the output that **ESPRESSO** will ultimately generate (see below) will be the required size of a case-control study with three times as many controls as cases (750/250) to produce the required power to detect the simulated effect. On the other hand, if you specify 1000 and 5000 it will generate the required size of a study with five times as many controls as cases (5000/1000). (2) **The absolute number of cases and controls.** The smaller you make the sample size, the faster the program will run, but the more variable the results will be (therefore requiring more simulations) and if you specify a number that is very small, rare combinations may not occur at all in a data set and this might generate misleading results, violate asymptotic assumptions, or it could even make the model fail altogether. As a rule of thumb, and given that for many real problems the required sample size will involve thousands of cases, it is probably wise to use at least 1,000 cases in your simulated sample. Default values are set to 2,000 cases and 8,000 controls (*i.e.* a design with four times as many controls as cases).

#6. Is there an interaction between the genetic determinant and the life-style determinant?

This is a binary indicator variable taking the value 1 if an interaction is to be included and 0 if not.

#7. Is the genetic model additive?

This is also a binary indicator variable taking the value 1 if the genetic effect is to be modelled as additive (*i.e.* using a covariate that is "the number of minor alleles", 0, 1 or 2) and 0 if it is to be modelled as a binary exposure.

#8. The population prevalence of disease

This parameter represents the prevalence of the disease in the general population on which the study, for which the power calculations are being undertaken, is to be based. In fact the estimated size requirement for a case-control study seems relatively robust to this parameter and a rough approximation is acceptable (though it appears to become more sensitive in the

presence of greater measurement error). If the true prevalence is very low, a very large number of individual subjects will have to be simulated before enough cases are generated and this will make the simulation process very slow. If the prevalence is less than 1 per 1,000 we often choose, therefore, to simulate assuming that the prevalence is 1 per 1,000. The default prevalence is set (completely arbitrarily) to 0.025.

#9. The Minor Allele Frequency (MAF)

This parameter represents the population prevalence of the rarer (minor) allele. The default prevalence is set (completely arbitrarily) to 0.3.

#10. The population prevalence of 'at risk' level of environmental factor

This parameter represents the prevalence of the 'at risk' life-style determinant in the general population on which the study is to be based. The default prevalence is set (completely arbitrarily) to 0.1.

#11. The odds ratio (OR) associated with the 'at risk' genetic variant

This is the odds ratio associated with a 1 unit increase in the genetic covariate (the extent to which the odds of disease is multiplied by being 'at risk' rather than 'not at risk' in a binary genetic model, or the multiplicative effect of each additional minor allele in an additive genetic model. The default OR is set (completely arbitrarily) to 1.5.

#12. The OR associated with the 'at risk' level of environmental factor

This is the odds ratio reflecting the ratio of the risk of developing the disease in subjects exposed to the 'at risk' level of the environmental determinant compared to those that are not exposed to this 'at risk' level. The default OR is set (completely arbitrarily) to 1.5.

#13. Interactive OR

In a binary model, this is the ratio of the odds ratio associated with having an 'at risk' genotype rather than a 'not at risk' genotype in subjects exposed to the 'at risk' lifestyle determinant compared to the same odds ratio in subjects not exposed to the 'at risk' lifestyle determinant. It should be noted that this is precisely the same as the ratio of the odds ratio associated with the 'at risk' lifestyle determinant in subjects exposed to the 'at risk' genotype compared to the same odds ratio in subjects not exposed to the 'at risk' genotype. If this odds ratio is 1.0 it implies that neither determinant influences the odds ratio associated with the other, and there is therefore '*no interaction*'. In an additive genetic model, it is the equivalent ratio of the odds associated with each additional minor allele. The default OR is set (completely arbitrarily) to 1.5.

#14. Baseline OR for subject on 95% population centile v 5% centile

This parameter reflects the heterogeneity in baseline disease risk (*i.e.* the heterogeneity in disease risk arising from determinants that have not been measured or have not been included in the model. Under the model currently implemented, it is assumed that the variation in baseline disease risk is normally distributed on the logistic scale. If this parameter is set to 10, the implication is that a 'high risk' subject (someone at the upper 95% centile of population risk) is, all else being equal, at 10 times the odds of developing disease compared to someone else who is at 'low risk' (at the lower 5% centile of population risk). This implies a $\log(10) = 2.3026$ difference on the $\log(\text{odds})$ scale. But the 5% and 95% centiles of a standard normal distribution occur at -1.645 and + 1.645 standard deviations. This means that, on the $\log(\text{odds})$ scale, 2.3026 corresponds to $(2 \times 1.645 = 3.29)$ standard deviations) and that the normal

distribution that provides the correct baseline heterogeneity in disease risk has a standard deviation of $2.3026/3.29 = 0.6999$. In essence, when parameter #14 is specified, a normally distributed random effect with a mean of 0 and a standard deviation of $\log(\text{parameter \#14})/3.29$ is generated.

The default value for this parameter is 10. In much of our earlier work we used 12.355 as the default setting. Although this may appear to be a surprising choice it corresponds to an approximate 20 fold ratio of risks between a very high risk subject on the 97.5% population centile and a very low risk subject on the 2.5% centile.

#15. P value defining statistical significance

This is totally up to the user, but it must take appropriate account of the related concepts of 'there being a very low prior probability that any one SNP will truly be associated with disease' and/or 'multiple testing'. We typically use $p < 10^{-4}$ for testing 'vague' candidate genes (candidature defined by vague biology, or location under a linkage peak), $p < 10^{-7}$ for genome wide association scans (GWASs), and we have tried using 10^{-10} for gene-gene interactions in a GWAS. For lifestyle determinants we typically use $p < 0.01$.

#16. Statistical Power

This is again up to the user. We prefer to design studies with a power of 80% to detect effects of true interest, but in the field of genetic epidemiology, 50% power is sometimes used.

#17. Sensitivity of genotype and #18. Specificity of genotype

The principal benefit of the *ESPRESSO* power calculator is that it enables power calculations to be carried out, taking *appropriate* account of measurement/assessment error in both outcome and explanatory variables. One of the important explanatory variables is the genotype.

A fundamental issue in constructing *ESPRESSO* was how best to define assessment error in a way that was meaningful to the user. For the purposes of the genetic determinant we felt that the most meaningful way was to relate it to the concept of r^2 - *i.e.* a measure of linkage disequilibrium (LD) that is, in effect, the squared Pearson correlation coefficient between the variables reflecting the alleles of two linked SNPs coded as 1s and 0s. That said, however, it is important to note that choosing to set the genotypic assessment error as being equivalent to, say, $r^2 = 0.8$ does not imply that we believe that genotypic measurement error is consequent *solely* on incomplete LD. Rather we choose to represent genotypic assessment error as if it were incomplete LD, because that representation may well be meaningful to users.

In practice, in order to keep the means of generating measurement error consistent across genetic and lifestyle explanatory variables, as well as outcome variables, the extent of measurement error is actually controlled by fixing the sensitivity and specificity of each variable (see supplementary materials, and the R code for *ESPRESSO*). So, r^2 is not specified directly: rather, one specifies the sensitivity and specificity that are required to generate an r^2 of the value desired for a SNP with a MAF that has been specified. There are therefore hyperlinks beside input parameters #17 and #18 leading to a program that generates the appropriate sensitivities and specificities that are required to generate the r^2 specified (given the specified MAF). The default values that are specified correspond to $r^2 = 0.8$, for a MAF of 0.3 (the default value for the MAF).

The sensitivity of a binary [1,0] variable is the proportion of true positives (which should therefore be coded 1) that are actually coded 1. Specificity is the proportion of true negatives (which should therefore be coded 0) that are actually coded 0.

#19. Sensitivity of environment and #20. Specificity of environment

Assessment/measurement error in an environmental determinant is determined by considering the reliability of a latent quantitative variable that is assumed to underlie the binary variable under consideration. This is an approach that is used widely in genetic epidemiology. It is sometimes called the latent threshold model. In effect, one assumes that there is a standardized normally distributed variable underlying the binary variable in question and if the value of this Gaussian variable exceeds a threshold T , the subject is “at risk” (binary variable = 1) and if it is less than T then the subject is “not at risk” (binary variable = 0). The value T is fixed at the value that corresponds to the correct prevalence of being “at risk” in the population under study. Measurement/assessment error may then be viewed as being quantified by the hypothetical reliability of that latent variable (see supplementary materials).

As before, hyperlinks beside input parameters #19 and #20 lead to a program that generates the appropriate sensitivities and specificities corresponding to the reliability that is required (given the prevalence of the at-risk life-style determinant).

The default sensitivity and specificity that are specified correspond to a reliability of 0.8 for an “at risk” prevalence of 0.1 (the default value of the prevalence of the ‘at risk’ level of the environmental determinant).

#21. Sensitivity of disease and #22. Specificity of disease

These are the sensitivity and specificity of the assessment of disease. Sometimes these are known directly. For example, the default values used in **ESPRESSO** correspond to the reported sensitivity and specificity of diagnosis of diabetes mellitus as diagnosed on the basis of a measured value of HbA1C (glycosylated haemoglobin) falling at or above the population 97.5% centile (*Rohlfing CL, Little RR, Wiedmeyer HM, England JD, Madsen R, Harris MI, et al. Use of GHb (HbA1c) in screening for undiagnosed diabetes in the U.S. population. Diabetes Care 2000;23(2):187-91*)

In consideration of parameters 14-19, the required value for any one may be obtained in whichever way may be preferred. For example, if it is known from pre-existing data that the true sensitivity and specificity of a particular environmental exposure measure are X and Y then clearly these values should be used.

RUNNING ESPRESSO INTERACTIVELY

1. Click on the **ESPRESSO** POWER CALCULATOR hypertext link
2. Specify values for all input parameters. Click on each box and replace the contents with your chosen value, or leave the default parameter as it is
3. Click on the hypertext links (“calculate”) beside parameters #17 and #18 to access the program allowing calculation of the sensitivity and specificity corresponding to a particular r^2 and a particular prevalence of the ‘at risk’ genotype prevalence. When you click on the link you will be asked to provide: (i) a random number seed (for an explanation of what you should enter, see input

parameter #2 for the main program); (ii) the prevalence of the 'at risk' genotype; (iii) the required r^2 . The r^2 program is iterative and may take a few minutes to run.

4. Click on the hypertext links ("calculate") beside parameters #19 and #20 to access the program allowing calculation of the sensitivity and specificity corresponding to the desired reliability of the latent quantitative variable underlying the environmental determinant. When you click on the link you will be asked to provide: (i) a random number seed (for an explanation of what you should enter, see input parameter #2 for the main program); (ii) the prevalence of the 'at risk' genotype; (iii) the required reliability.
5. Once all input parameters are specified to your satisfaction, click the "Run Calculation" button at the bottom. A new window opens up with a spinning (busy) icon and the program prints a line of output after every 5th simulation so you can be reassured that it is working. Using the default setting, 100 simulations take just under a minute to run and 500 simulations take approximately 3.5 minutes.

RUNNING THE PROGRAM IN BATCH

Download the [ESPRESSO.CC.batch.03Aug2008.R.rtf](#) file as described in the document ESPRESSO.WEBSITE.INFORMATION.03Aug2008.doc.

Also download the exemplar batch control file

[ESPRESSO.CC.batch.example.03Aug2008.control.csv](#) and the illustrative output file that this generates

[ESPRESSO.CC.batch.example.03Aug2008.output.csv](#).

Once you are happy with how it works, you can specify and run as many simulation scenarios as you wish, with output going to a batch output file.

SCREEN OUTPUT

When the program has run, one obtains the following output directly to the screen (in the window in which *ESPRESSO* is running).

1. It confirms what seed you set, how many simulations you ran, and how many cases and controls you actually simulated.
2. ***It tells you the number of cases and number of controls you would require in a design (with the number of cases and controls in the same ratio as they are in the data set you actually simulated) in order to be able to detect: (i) a genetic determinant; (ii) an environmental determinant; (iii) a gene-environment interaction of the sizes simulated under the specific conditions of the scenario as simulated (at the specified level of statistical power).

OUTPUT TO A FILE

A more comprehensive and permanent version of the output may be obtained by clicking on the download hypertext links at the top of the output window. The file produced contains a header line and a single line of output representing input and output parameters for each specified scenario. This is precisely equivalent to the file generated by the batch version of *ESPRESSO*.

The file is an Excel-like dataset in international .csv format. This means that it is semicolon (;) delimited rather than comma (,) delimited. If your computer is not set up

for this format (*this is a function that has to be set in the control panel, not in Excel*), the easiest way to read the file into Excel is to rename it with a .txt rather than a .csv extension and then to open it as a text file in Excel. Choose 'Delimited' format and then select semi-colon as the separator.

OUTPUT FORMAT

The file contains the following elements:

Columns A-V repeat the 22 input parameters

Column W repeats the scenario ID again

Columns X-AC contain the number of cases and controls required (see above^{***}) to detect the genetic and environmental main effects (and the interaction, if specified). If an interaction is not specified, the interaction columns contain missing values (NAs). If a main effects (only) model is required, you should fit a model without an interaction (*i.e.* input parameter #6 = 0) rather than specifying a model WITH an interaction (*i.e.* input parameter #6 = 1) and stating that the interaction has no effect (input parameter #13 = 1.0). These two models are NOT equivalent.

Columns AD-AF contain the model-based power for detecting the genetic, environmental and interaction effects given the number of cases and controls that were *actually simulated*. Columns AG-AI contain the equivalent empirical power. The difference between the model-based and empirical power is described in the document containing supplementary methods. It is explained that it is actually model-based power that *ESPRESSO* uses as the basis for calculating the required number of cases and controls in columns X-AC.

Column AJ repeats the scenario ID

Columns AK-AM contain the estimated ORs for the genetic, environmental and interactive terms. On theoretical grounds these would be expected to be shrunk relative to the simulated values (*Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. Stat Med 1992;11(14-15):1825-39.*)

Columns AN-AP contain the standard errors of the estimates in columns AK-AM

Columns AQ-BA contain empirical estimates of key features of the simulated data. These provide an approximate check that the simulated sensitivity and specificity of the alleles, environmental determinant and disease, and the r^2 reflecting assessment error in the alleles are all appropriate given the scenario intended. It should be noted that because the errors in the disease status are specified *before* sampling, the estimated sensitivity and specificity of the disease status are distorted. Specifically, because the sampling fraction of controls is normally much lower than that of cases, assessment errors in subjects that are truly diseased (that convert them from true cases to false controls) are under-sampled resulting in an overestimated sensitivity. In contrast, assessment errors in subjects that are truly disease-free (that convert them from true controls to false cases) are over-sampled resulting in an underestimated specificity.