USER MANUAL for REGENT:
Risk Estimation for Genetic and Environmental Traits

Daniel JM Crouch[1], Graham HM Goddard[1] & Cathryn M Lewis[1,2]
[1]Department of Medical and Molecular Genetics, [2]MRC Centre for Social, Genetic & Developmental Psychiatry, King's College London

email: daniel.crouch@kcl.ac.uk, cathryn.lewis@kcl.ac.uk

web:
http://www.kcl.ac.uk/medicine/research/divisions/gmm/sections/clusters/bse/lewis/index.aspx

CRAN: http://cran.r-project.org/web/packages/REGENT/index.html

Contents

1. Overview

This document contains instructions to use the REGENT R package for genetic risk categorisation and prediction. We recommend that new users read this document. We have taken care to make it understandable for users with a non-statistical background. Arguments for the REGENT functions (settings/options) are written in italics, and key words are highlighted in bold.

REGENT creates statistically valid risk categories based on the distribution of genetic and environmental risks in the population. Patients or subjects can then be fed into this model to produce a personal risk category. Details of the concepts and methods behind REGENT can be found in the following two papers

- Crouch, Goddard & Lewis, REGENT: a risk assessment and classification algorithm for genetic and environmental factors (2011)
- Goddard and Lewis: Risk categorisation for complex disorders according to genotype relative risk and precision in parameter estimates, Genetic Epidemiology (2010) 34(6):624-32.

In short, REGENT comprises two functions REGENTmodel and REGENTpredict, which are run in the R programming environment. REGENTmodel uses summary information on genetic and environmental risk factors to calculate the population distribution of risk. To predict risk for specific individuals, for example patients, their genotype and environmental risk factor data is fed into REGENTpredict together with the output from REGENTmodel. Individual-level output on absolute disease risk, relative disease risk, and risk categorisation (low, average, elevated or high) is provided.

2. What you will need

REGENT is simple to use, but the input data require careful consideration. Any figures entered for research, clinical or actuarial purposes must be statistically justified either by original research or by the scientific literature.

Users must also take care to provide figures which are not only statistically sound but appropriate for their population of interest; environmental factors, genetics and disease prevalence can be markedly different across ethnic and geographic groups.

Users should take the best estimates available – those from studies with the largest sample sizes, or if possible from meta-analyses.

The required data are listed below according to the type of analysis being performed. If known, these should all be available from epidemiological or genetic epidemiological

research papers, or at least attainable through simple calculations based on the reported figures.

We recommend rounding input figures to 2 decimal places. Additional decimal places imply an exaggerated precision.

Input data for disease prevalence

For either genetic or environmental analysis, the user must have an estimate for the **coefficient of variation (CV)** for disease prevalence (a measure of the uncertainty in the prevalence estimate). This is not usually reported directly in the literature but can easily be calculated by hand:

$$\text{CV} = \frac{\text{Standard deviation of prevalence}}{\text{prevalence}}$$

Ideally, both prevalence and the standard deviation for the prevalence can be found in the literature[1]. If you do not have these, then CV can be estimated from a single study of *n* individuals, *d* of who are affected by the disease, as:

$$\text{CV} = \sqrt{(1/d - 1/n)}$$

REGENT has a default setting of *cv*=0.05, but it is important to consider whether this is suitable your disease of interest, as the value can significantly influence the results. Higher values indicate that disease prevalence is known with less precision.

Input data for genetic analysis

For each SNP you wish to analyse, you will require:

1. A **name**, for example an rsID. You can make one up if you wish, but each SNP should have a unique name.
2. A **risk ratio**.
    - This is a figure which describes the risk of developing a disease for an individual with one risk allele, relative to an individual with no risk alleles.
    - A risk ratio of 2 means that a single allele doubles the risk of disease.
    - Risk ratios below 1 mean that the allele protects against disease.

[1] Sometimes a standard deviation will be unavailable, but 95% confidence intervals are. You can easily calculate it from the 95% confidence intervals as follows: SD=(UCI-prevalence)/1.96, where UCI is the upper 95% confidence interval.

- The risk ratio for 2 alleles is assumed to be the square of the single allele risk, but you may also supply separate risks for the two genotypes – details of this in section 3.
- Risks are usually reported as odds ratios, relative risks or genotype relative risks, depending on the design of the original study, and you may take any of these as the risk ratio.

3 The number of **cases** that were used to calculate the risk ratio
4 The number of **controls** that were used to calculate the risk ratio
5 An **allele frequency** for the risk allele in the population - a number between 0 and 1.
6 **Coefficient of variation** (see above)

Note that the SNPs alleles (e.g. A, C) are not required, but the allele frequency and risk ratios must tie up with each other

Input data for environmental analysis

For each environmental factor you wish to analyse, you will require:

1 A **name**, e.g. Smoking. Each factor should have a unique name.
2 One or more **risk ratios**
   - This is a figure which describes the risk of developing a disease for an individual with exposure to the factor, relative to an individual with no exposure.
   - A risk ratio of 2 means that exposure doubles the risk of disease.
   - Risk ratios below 1 mean that the factor protects against disease.
3 The **standard error(s)** for the log of the risk ratios – see below
4 One or more **exposure frequencies** to the risk level(s) in the population - a number between 0 and 1.
5 **Coefficient of variation** (see above)

You may provide multiple exposure levels for risk factors. For example, smoking and ex-smoking could be analysed as a two-level risk factor, both levels compared to the baseline risk for non-smokers. These share the same **name**, but all other variables must be provided separately for each level (risk ratios, standard errors, exposure frequencies, CV). Details of how to do this are in section 3.

The **standard error** for the log of the risk ratio is not usually reported in the literature, but can be easily calculated from the 95% confidence intervals as

$$SE = [\log(UCI)-\log(RR)]/1.96$$

where RR is the risk ratio estimate (e.g. an odds ratio) and UCI is the upper 95% confidence interval, which is usually reported. log() is the natural log function[2]. This should provide the same answer as

$$SE=[\log(UCI)-\log(LCI)]/3.92$$

where LCI is the lower 95% confidence interval. If it does not, there is probably a small rounding error in the reported figures.

Input data for REGENTpredict

For each individual to be analysed, you must have a **genotype** for each SNP in the model, and an **exposure level** for each environmental factor.

Full genotype and environmental data must be provided to calculate individual-level risks, as REGENTpredict does not allow for missing data. There are several methods to deal with an individual with missing risk factor data.

1. The best solution is to use REGENTmodel to create a new model limited to the SNPs and factors which are present.
2. You could replace the missing data item with each possible value in turn (e.g. three genotypes) and use REGENTpredict to calculate the risk at each value. The risks can be combined, weighting by the probability of the genotype, so that the final estimate averages over all possible values of the missing data item.
3. Genotypes may be imputed if you have data present on flanking SNPs in high linkage disequilibrium (e.g. using MACH, BEAGLE or IMPUTE)
4. For a data set with many individuals present, multiple imputation methods can be used possible to create representations of the full data set. REGENTpredict is then run on each representation and standard statistical methods are used to combine results across the datasets generated.

We advise caution when high risk variables are missing, as these have the largest influence on results.

If you are interested in the absolute risk of disease, you must have a good estimate of its **prevalence**.

[2] The natural log function can be performed by most pocket calculators, but is sometimes named confusingly: 'log' can sometimes refer to the log base 10 function. Check that you are calculating the natural log by taking the log of 2.718282. If the answer is 1, you have the correct function. The log() function in R calculates the natural log by default.

3. Running REGENT

Step 1: Create input files for REGENTmodel

Example input files are in …/R/library/REGENT/data.

If you wish to model SNPs, create a text file[3] in the following format in the directory you wish to work in:

SNP	MAF	RR	Ncase	Ncontrol
rs11209026	0.932	2.66	1639	1808
rs3792109	0.529	1.34	6333	15056
rs11742570	0.606	1.33	6333	15056

- The first row should be the same as in the examples, and the rest changed to match your entries.
- There should be as many lines as your number of SNPs, plus the first row.
- You can call this file whatever you want, but it is sensible to put a ".txt" as the file extension.
- Extraneous columns may be provided and will simply be ignored by the program.
- Names must not have spaces

The entries in this table, and how to find them, are explained in section 2. The abbreviations are as follows: SNP refers to the **name** of the SNP, MAF to the **allele frequency** of the risk (can be protective) allele, Ncase to the number of **cases** and Ncontrol as the number of **controls.**

If you wish to specify risk ratios separately for single risk allele and double risk allele genotypes (heterozygotes and homozygotes), the file should look like this:

SNP	MAF	RR_het	RR_hom	Ncase	Ncontrol
rs11209026	0.932	2.66	7.08	1639	1808
rs3792109	0.529	1.34	1.80	6333	15056
rs11742570	0.606	1.33	1.77	6333	15056

Note that files are shown formatted for clarity in this document, but this is not necessary for data input files. Creating a file for environmental inputs is similar and the same points apply, but the format looks like this:

Factor	Exposed	RR	SE
Smoking	0.27	2.02	0.103

[3] In windows this can be done quickly by right clicking in the directory, then going to New->Text Document in the menu. In Mac OS X, the default format used by TextEdit is .rtf, but you can convert to plain text by going to Format->Make plain text.

Factor is the **name** of the risk factor, Exposed refers to the **exposure frequency**, RR the **risk ratio** and SE the **standard error of the log of the risk ratio**.

If you want multilevel factors, the format is slightly different:

Factor	Exposed1	Exposed2	RR1	RR2	SE1	SE2
Smoking	0.17	0.1	1.5	2.02	0.103	0.105

Note that all the variables apart from Factor here are numeric.

Step 2: Execute commands

Open R. The first thing to do is set your working directory:

```
setwd("…")
```

Type the above command into the interface and, where the dotted lines are, type the file path to the directory which your input files are in. R needs forward slashes rather than backslashes.

If you have not already done so, install REGENT:

```
install.packages("REGENT")
```

A menu will appear asking you to select a server based on location. Select the one nearest to you and click 'ok'.

Next, load the REGENT package:

```
library(REGENT)
```

Then, to run REGENTmodel, enter the following command, filling in the relevant information where the dotted lines are.

```
x=REGENT.model(AnalysisName="…",LocusFile="…",EnvFile="…",cv=…)
```

- AnalysisName will be what your output files are named after. It should be something which describes the run you are performing. If you perform subsequent runs with the same *AnalysisName* setting, the original output files will be overwritten.
- In the dotted lines after LocusFile, fill in the name of the file you created containing the SNP data including the file extension (usually .txt).
- In the dotted lines after EnvFile, fill in the name of the file you created containing the environmental data

- after cv enter a number for the coefficient of variation[4].
- You can put any character you want[5] in place of x – this variable will store parts of the output.

If you did not prepare an environment file, then leave out the *EnvFile* argument:

x=REGENT.model(AnalysisName="…",LocusFile="…",cv=…)

Likewise, if you did not prepare a SNP file, leave out the *LocusFile* argument.

After you press enter, R will feed updates on the progress to the user interface. When the analysis is completed, look in the working directory to find the 4 output files, named with suffixes '.txt', '_RRdist.txt', '_RRdistCol.TIF' and '_RRdistGrey.TIF' after the prefix specified by *AnalysisName*.

Step 3: Using REGENTpredict

Create another input file in the following format:

rs11209026	rs3792109		rs11742570		Smoking
Person_A	0	0	0	0	
Person_B	0	2	2	0	
Person_C	2	0	0	1	
Person_D	2	2	2	1	

- The variable names at the top of the file must have the same names as were provided to risk model. It is not enough for the order to be the same (it may be different), as the program matches by name.
- You may name your individuals as you please as long as there are no spaces, but it is best to have a unique name for each person
- For SNPs, the number gives the number of copies of the allele (0, 1, or 2) for which RR and MAF were provided for in REGENTmodel. It is essential that alleles are coded consistently across the REGENTmodel and REGENTpredict analyses.
- For environmental factors, the number describes the exposure level of the individual (0, 1 for a binary level factor; 0, 1, 2 for a three-level factor, where 0 is the baseline level).

Run REGENTpredict as follows:

y=REGENT.predict(AnalysisName="…",ind="…",model="…")

- Provide *ind* with the name of the file you created with the individual data
- model may either be the variable x from above if you are in the same R session, or the name of the file '*AnalysisName*.txt' created by REGENTmodel.

[4] *cv* has a default of 0.05. If you happy with a cv of 0.05 you may omit this argument
[5] R variables must start with a letter rather than a number

- *AnalysisName* may have the same setting as REGENTmodel without overwriting files.

The single output file is called '*AnalysisName*_Predictions.txt'

4. Interpreting the results

REGENTmodel

The colour figure produced by REGENTmodel, in the file *AnalysisName*_RRDistCol.TIF, is the key to understanding the landscape of risks in the population. The Y-axis is the risk to an individual, relative to the individual with the average risk – denoted by the grey horizontal line at 1. The X-axis is the percentage of the population with a risk equal to or lower than the value on the Y axis.

The risk categories are indicated by colour. If the ability to categorise individuals is low, then many individuals will fall into the 'average' category, coloured pale blue. We cannot say with statistical confidence that individuals this category have a different relative risk to the average member of the population (the grey line), even though our best estimate of their relative risk may be different from 1.

Good discrimination will cause the red, yellow and green sections to increase in size, meaning that you can say with statistical confidence that these individuals have a different risk to the average member of the population. The high risk category, coloured red, consists of individuals who we can be confident have greater risk than those in the elevated (yellow) category. The proportion of the population in each risk category can also be found in the main output file (*AnalysisName*.txt).

The level of discrimination is driven by the quality of the risk estimates for the SNPs and environmental factors, and by their size and frequency. All of these factors interplay. Discrimination can be improved by increasing sample sizes and discovering additional risk factors.

The additional 2 text output files are intended for advanced users.
'*AnalysisName*_RRdist.txt' is for recreating graphs of your own based on the raw risk distribution output, whilst '*AnalysisName*.txt' contains details of the model, all input parameter settings and input data.

REGENTpredict

For each individual, REGENTpredict gives:

1. **Absolute risk.** This is an estimate of the probability that the individual will develop disease. It depends on the value of disease *prevalence* and as such should be treated with some care.

2. **Relative risk.** This is the risk that an individual will develop disease relative to the average member of the population, who has a relative risk of 1. It includes the risk conferred by genetic and environmental factors.
3. **Risk category.** Describes whether the risk for this individual can be considered statistically higher or lower than an average member of the population.

Most users will be primarily interested in the **Relative risk** and **Risk category.**

For advanced users, REGENTpredict also provides:

1. Lower confidence interval for the relative risk – according to the value provided for *alpha,* a variable for both functions, which determines the width of the confidence interval (default 0.95).
2. Upper confidence interval for the relative risk – according to the value provided for *alpha*
3. Borderline status. If 'yes' the individual is on the borderline of the given risk category. Changing *alpha* by ±1% would reclassify the individual into higher/lower category.

5. FAQs

Q: REGENTmodel is taking a long time to run. Is this normal?
A: REGENTmodel takes just a few seconds to run with a few risk factors, but as the number exceeds 5 it can begin to take substantially longer. With 10 SNPs it takes approximately 15 minutes, and with ~70 SNPs it takes approximately 8 hours. REGENTmodel does not need to be run often since all output is saved to file for REGENTpredict to utilise whenever necessary.

Q: Are there any ways to decrease run time?
A: Run time can be decreased by using the *Block* argument in REGENTmodel. The default is 100, but increasing this value will hold more genotypes in memory, allowing more operations to be vectorised which speeds up the calculations. However if set too high you can run out of memory. Somewhere between 100 and 300 is about right for a personal computing system with 2GB of RAM. If you are using a high performance computing cluster, *Block* can be substantially increased, for example up to 10,000, which should speed things up significantly. Note that to use more than 4GB of RAM, R-64bit must be used.

Q: I accidently closed R between using REGENTmodel and REGENTpredict. Have I lost my model?
A: No, the model can be read in from the main output file, named according to the setting *AnalysisName*.

Q: I get different risk category figures when I repeat my REGENTmodel analysis on the same data – is this a mistake?
A: No, the risk categories are based on simulations and so the results will be slightly different each time. However the variability in the numbers should be low. It is best to run analyses at least twice to check that your first run was not based on an unusual simulation. The numbers of simulations (*indsim)* can also be increased from the default value of 100,000 (e.g. to 1,000,000).

Q: Can I model interactions between variables?
A: REGENT assumes independence between all genetic and environmental factors, but it is possible to model interactions by entering all the possible combinations of two (or more) variables, along with their frequencies and risks, as a multilevel 'environmental factor'.

Q: Will REGENT model missing data?
A: No, there are no plans to incorporate missing data modelling. For suggestions on how to deal with missingness please see 'input data for REGENTpredict' under section 2 of this document.

Q: Why is there an arrow next to the output graphs?
A: Diseases with some high risk factors tend to have a small proportion of the population with very high risks. By default, the Y-axis is truncated at 5, as most users will be more interested in the common variation. The arrow is there to indicate that some risks extend higher than the limit of the Y-axis. The limit can be changed by altering the *PlotMax* argument.

Q: I get this warning message: 'Individual (s) … move to higher risk category when homozygotes at SNP(s) are changed to heterozygotes'. What does it mean?
A: Sometimes the confidence interval for a homozygous risk allele genotype is wider than that for the heterozygote. This can cause homozygous individuals to be classified into a lower risk category even though this goes somewhat against the biology. Some users may wish to consider individuals with 2 risk alleles as having equal or higher risk than those with 1 risk allele by definition, even though the confidence interval may be wider for the former. If so, we recommend running the individuals again with those SNPs listed in the warning message set to heterozygotes, and take the resulting risk categories as your answer. However, all other results (absolute risk, genotype relative risk, upper and lower confidence intervals) must be based on the run using the individual's true genotype.

Q: I get the warning message 'incomplete final line found by readtableheader'
A: Add an empty line to the end of your input files (place the cursor to the right of the final character and press enter). This warning message is nothing to worry about and your results should not be affected.