Multifactorial Disease Risk Calculator: Web-based risk prediction for multifactorial disease pedigrees

Article Type – Brief Communication

Running Title

Risk prediction for multifactorial disease pedigrees

Authors

Desmond D Campbell[1,2,*], Yiming Li[2] and Pak C Sham[2,*]

Affiliations

1 Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK

2 Department of Psychiatry and Centre for Genomic Sciences, University of Hong Kong, Hong Kong SAR

*To whom correspondence should be addressed.

Contact

Desmond Campbell

Room 239, BHF Cardiovascular Research Centre

Institute of Cardiovascular and Medical Sciences, University of Glasgow,

126 University Place, Glasgow, G12 8TA, United Kingdom

Tel +44 141 330 7615

Desmond.Campbell@glasgow.ac.uk

Abstract

Multifactorial Disease Risk Calculator is a web tool automating construction of multifactorial disease models from epidemiological findings. It provides a user friendly interface for doing this, extending a reported methodology based on a liability-threshold model in several ways. Multifactorial disease models incorporating all the following features in combination are handled: quantitative risk factors (including polygenic scores), categorical risk factors (including major genetic risk loci), stratified age of onset curves, and the partition of the population variance in disease liability into genetic, shared and unique environment effects. It allows the application of such models to disease pedigrees. Pedigree related outputs are (i) the disease pedigree liability distribution, (ii) individual disease risk for pedigree members, and (iii) *n* year disease risk for pedigree members. Risk prediction for each pedigree member is based on using the disease model to appropriately weigh evidence on disease risk available from personal attributes and family history. Evidence is used to construct the disease pedigree liability distribution. From this predictions of lifetime and *n* year risk can be calculated. Example disease models and pedigrees are provided at the website and are used in accompanying tutorials to illustrate the features available. A command line version is available. Website:
http://grass.cgs.hku.hk:3838/mdrc/current

Key words

Disease pedigree, Liability Threshold Model, Risk Prediction, Software

Introduction

Individual risk prediction for multifactorial disease is an important goal. Several websites addressing this, based on family history, have been developed (Facio et al. 2010; Yoon et al. 2009). Combining family history with known risk factors can improve risk prediction, and family history and polygenic score have been found to contribute to risk prediction in a complimentary manner (Do et al. 2012). Several similar risk prediction methodologies that combine family history with known risk factors have been developed (Campbell et al. 2010; Ruderfer, Korn, and Purcell 2010; So et al. 2011). Here we describe a web tool based on one of these. Although predictive power for multifactorial diseases will probably remain modest, decreasing assay costs and improvements in polygenic score accuracy, along with more accurate disease models, will increase the relevance of such predictions for health related decision making. Sources of bias and inaccuracy in such predictions are reviewed in (Do et al. 2012). By combining all available evidence risk predictions may be regarded as a data reduction step and in this regard may have utility in medical research (we do not develop that idea here).

For many multifactorial diseases, a model can be specified based on existing research findings. Estimates for the relative contributions of genetic, shared environmental and unique environmental effects are often available from twin and adoption studies. Lifetime risk estimates are available from family and epidemiology studies. Lifetime risk often differs across population strata, the variable defining the strata (e.g. sex) being a risk factor for the disease. Many multifactorial diseases are not congenital. How onset relates to age is a disease characteristic, and is often described by an age of onset curve. Age of onset curves may also vary across population strata. Personal attributes such as age, sex, and polygenic score, along with family history for the disease, can be combined with a disease model developed from previous research findings, to predict individual risk. A method for doing this, based on a liability threshold model for multifactorial disease, has been developed and implemented in software (Campbell et al. 2010). This builds a prior disease liability distribution for the pedigree based on the pedigree structure. From this, a posterior distribution is generated, by conditioning on pedigree members' personal attributes including affection status. Pedigree member risks are predicted from this posterior liability distribution.

Features

The web tool extends the (Campbell et al. 2010) methodology in several ways:

1. Derivation of the appropriate disease model from epidemiological findings is difficult when categorical risk factors are involved. A worked example is given in (Campbell et al. 2010), where a disease model is derived that accounts for Depression being twice as prevalent in females as males but this is a simple case. The web tool automates model derivation. How that is done is described below.
2. Quantitative risk factors (e.g. BMI and polygenic score) can now be incorporated into the disease model. The risk factor's impact on disease liability and how that is partitioned across genetic and environmental components has to be specified.
3. Multiple categorical and/or quantitative risk factors can be modelled, and these can be correlated. How risk factors covary, for each genetic and environmental component of liability, must be specified.
4. Shared environmental effect can now be modelled (in addition to additive genetics and non-shared environment).
5. Expressed proportion of lifetime risk is a personal attribute of relevance for non-congenital diseases. It is the probability of a person having already manifested the disease (given their age etc.) given that they would ultimately do so if they lived long enough, and is a measure of the right censoring of affection status information. For each person, expressed proportion of lifetime risk is interpolated from their age and their age of onset curve. Previously it was assigned by age category.

6. Pedigree members' n year risks can now be predicted. Inclusion criteria for several public screening programs are based on this.
7. Pedigrees containing twin relationships can now be specified
8. The web tool also performs more rigorous checks of the pedigree's and the disease model's validity.

Categorical risk factors are incorporated into a multifactorial disease model as follows. For each risk factor category, the user specifies its population frequency, and relative risk (relative to an arbitrarily chosen reference category). How the liability explained by the risk factor is split up over the genetic and environmental components of liability must also be specified. The disease is modelled using a liability threshold model in which the population liability distribution is a mixture distribution of risk factor strata liability distributions. The mixture proportions are the frequencies of the risk factor strata. The liability distribution within each stratum is assumed to be Gaussian. The means of these liability distributions are allowed to differ across strata, but their variances are constrained to be equal. The population liability distribution is further constrained to have a mean of 0 and variance of 1. The per strata means, the within-stratum variance, and the appropriate critical threshold can then be found by solving a set of non-linear simultaneous equations.

To allow an *n* category risk factor to be incorporated into the disease model, the user must supply

- K = Disease lifetime risk
- m_i = the relative risk for ith non-reference stratum (of which there are $n-1$).These m_i being the risks relative to an arbitrarily chosen reference stratum
- f_i = the frequency of the ith non-reference stratum

Given these inputs, the frequency and lifetime risk for all strata can be directly calculated. It is then possible to specify *n* simultaneous equations in *n* unknowns (for the purposes of this exposition the reference stratum is the *n*th stratum). These equations are

$$\int_{T}^{\infty} \varphi(\mu_i, \sigma^2)\, dx = m_i \int_{T}^{\infty} \varphi(\mu_n, \sigma^2)\, dx$$

$$r_n = \int_{T}^{\infty} \varphi(\mu_n, \sigma^2)\, dx$$

where

- φ is the Gaussian probability density function
- $\sigma^2 = 1 - \sum_{i=1}^{n}(f_i {\mu_i}^2)$ is the intra stratum liability variance

the *n* unknowns being

- *T* = critical threshold of liability
- μ_i= liability mean for ith non-reference stratum

These simultaneous equations are solved iteratively. Based on experimentation, we settled on using different optimisation methods depending on the number of risk factor categories

- n=2 – BFGS (Broyden et al. 1970)
- n>2 – Nelder-Mead (Nelder and Mead 1965)

The optimisation algorithms iteratively improve their guess of the unknown parameters' values by measuring the discrepancy between the target strata risks (previously calculated), and the strata risks that would follow given the current guess. The discrepancy function used is

$$d = \begin{array}{ll} \boldsymbol{f}'\left[(\boldsymbol{r} - \tilde{\boldsymbol{r}})^{\boldsymbol{2}}\right] + \left(\frac{1}{n}\right)\mathbf{1}'\left[(\boldsymbol{r} - \tilde{\boldsymbol{r}})^{\boldsymbol{2}}\right] & , if\ \sigma^2 \geq 0 \\ 2 - \sigma^2 & , if\ \sigma^2 < 0 \end{array}$$

where

- $\boldsymbol{r}$ = the target per strata lifetime risks
- $\tilde{\boldsymbol{r}}$ = the per strata lifetime risks according to the current guess

Considering this function when $\sigma^2 \geq 0$, the idea here is that if one of the two terms is satisfied then the other term will dominate. The first term is the population average of the squared Euclidean distance of $\tilde{\boldsymbol{r}}$ from the target. Optimisation using this term alone as the discrepancy performed poorly for risk factors that contained one or more rare categories. The extra term up-weighs rare categories guarding against the possibility of a good population fit coming at the price of a disastrous fit for rare strata.

It is possible for the unknowns to take values that imply the liability variance explained by the risk factor strata exceeds 1. This is invalid given that the population liability variance constraint, and manifests as σ^2 being negative. This scenario must be handled as there is nothing preventing such guesses being made during optimisation. The rationale in this scenario is as follows. When $\sigma^2 \geq 0$ then $d < 2$, whereas when $\sigma^2 < 0$ then $d > 2$. Consequently d for an invalid guess is guaranteed to be worse than d for any valid guess. Also the more negative σ^2 is, the worse d is. Thus the discrepancy function gradient points in the direction of the valid parameter space.

Example disease models and pedigrees are available via the website. These are used in accompanying tutorials to explain the features available. Disease models have been pre-specified for several common multifactorial diseases. Text specifying the disease model can be uploaded, edited and saved via the web interface. The text is monitored, and in response to edits, an attempt is made to build the disease model. If successful the constructed disease model is reported, otherwise an appropriate error is reported. Pedigree information can be uploaded, edited and saved via the web interface. The pedigree information is monitored, and in response to edits, the validity of pedigree information is checked, and detected errors reported. If valid then the corresponding pedigree diagram is generated and displayed. Pedigree information is specified in the most commonly used ped/linkage format. This is a flat file format with one line per pedigree member. The first four variables of each line (pedigree id, person id, father id and mother id) define pedigree structure. These along with the following two variables (sex and affection status) allow classic disease pedigree diagrams to be constructed. Further variables are optional, although in the context of this website several such have special meanings (e.g. age is relevant for non-congenital disease models). Pedigrees are modified by editing ped-formatted text. Some pedigree software allows graphical editing of pedigrees. We don't find this a great advantage; pedigree diagram generation mitigates against the difficulty of visualising the pedigree information, and the website's main purpose is risk prediction not pedigree specification. Comment lines are supported. This allows persons to be quickly removed from and restored to the pedigree. Once a valid disease model and pedigree have been specified, individual risk for the pedigree members can be calculated.

A command line version of the risk prediction program is available that provides some extra functionality. In particular it writes to file a sample dataset drawn from the disease pedigree's posterior liability distribution. An idea of risk prediction precision can be obtained by running the risk prediction repeatedly. The command line program allows a more formal estimation of risk prediction precision. Also downloadable is documentation, in particular regarding how categorical risk factors are handled, and the testing done to validate the software.

Implementation

The program is written in R (Team 2016) and C++. The website was built using R package *shiny* (Chang et al. 2015). R package *kinship2* is responsible for much of the pedigree related functionality (Sinnwell, Therneau, and Schaid 2014). A Gibbs sampler (written in C++ for speed) is incorporated using R package *Rcpp* (Eddelbuettel and François 2011).

Conclusion

Multifactorial Disease Risk Calculator is a web tool automating construction of multifactorial disease models from epidemiological findings. It allows the application of such models to disease pedigrees. Pedigree related outputs are (i) the disease pedigree liability distribution, (ii) individual disease risk for pedigree members, and (iii) n year disease risk for pedigree members. Availability: http://grass.cgs.hku.hk:3838/mdrc/current

Acknowledgments

This work was supported by the European Community's Seventh Framework Programme grant HEALTH-F2-2010-241909 (Project EU-GEI).

Conflicts of Interest

None

References

Broyden, C. G., R. Fletcher, D. Goldfarb, and Shanno, D. F. 1970. 'J. Inst'. *Math. Appl* 6: 76.

Campbell, Desmond D., Pak C. Sham, Jo Knight, Harvey Wickham, and Sabine Landau. 2010. 'Software for Generating Liability Distributions for Pedigrees Conditional on Their Observed Disease States and Covariates'. *Genetic Epidemiology* 34 (2): 159–70. doi:10.1002/gepi.20446.

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2015. *Shiny: Web Application Framework for R*. http://CRAN.R-project.org/package=shiny.

Do, Chuong B., David A. Hinds, Uta Francke, and Nicholas Eriksson. 2012. 'Comparison of Family History and SNPs for Predicting Risk of Complex Disease'. *PLoS Genetics* 8 (10). doi:10.1371/journal.pgen.1002973.

Eddelbuettel, Dirk, and Romain François. 2011. 'Rcpp: Seamless R and C++ Integration'. *Journal of Statistical Software* 40 (8): 1–18.

Facio, Flavia M., W. Gregory Feero, Amy Linn, Neal Oden, Kandamurugu Manickam, and Leslie G. Biesecker. 2010. 'Validation of My Family Health Portrait for Six Common Heritable Conditions'. *Genetics in Medicine* 12 (6): 370–75. doi:10.1097/GIM.0b013e3181e15bd5.

Nelder, J. A., and R. Mead. 1965. 'A Simplex Method for Function Minimization'. *The Computer Journal* 7 (4): 308–13. doi:10.1093/comjnl/7.4.308.

Ruderfer, Douglas M, Joshua Korn, and Shaun M Purcell. 2010. 'Family-Based Genetic Risk Prediction of Multifactorial Disease'. *Genome Medicine* 2 (1): 2. doi:10.1186/gm123.

Sinnwell, Jason P., Terry M. Therneau, and Daniel J. Schaid. 2014. 'The kinship2 R Package for Pedigree Data'. *Human Heredity* 78 (2): 91–93. doi:10.1159/000363105.

So, Hon-Cheong, Johnny S.H. Kwan, Stacey S. Cherny, and Pak C. Sham. 2011. 'Risk Prediction of Complex Diseases from Family History and Known Susceptibility Loci, with Applications for Cancer Screening'. *American Journal of Human Genetics* 88 (5): 548–65. doi:10.1016/j.ajhg.2011.04.001.

Team, R Core. 2016. *R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2014*.

Yoon, Paula W., Maren T. Scheuner, Cynthia Jorgensen, and Muin J. Khoury. 2009. 'Developing Family Healthware, a Family History Screening Tool to Prevent Common Chronic Diseases'. *Preventing Chronic Disease* 6 (1): A33.