Lietuvos matematikos rinkinys
Proc. of the Lithuanian Mathematical Society, Ser. A
DOI: 10.15388/LMR.A.2015.19

ISSN 0132-2818
Vol. **56**, 2015
pages 107–112

Analysis of genetic risk assessment methods

**Vytautas Tiešis, Algirdas Lančinskas,
Virginijus Marcinkevičius**

Institute of Mathematics and Informatics, Vilnius University
Akademijos 4, LT-08663 Vilnius
E-mail: vytautas.tiesis@mii.vu.lt, algirdas.lancinskas@mii.vu.lt
E-mail: virginijus.marcinkevicius@mii.vu.lt

Abstract. Chronic non-communicable diseases are caused by a combination of multilocus genetic risk factors. The genetic risk assessment companies, e.g. Navigenics and 23andMe, calculate a lifetime risk of a disease by the use of strong assumptions on the total impact of the multiple SNPs genotype. The object of the paper is to compare such risk assessment methods. The theoretical disease model that describes both environmental and genetic factors has been used for evaluation of assessment methods. The system of nonlinear equations for tuning model's parameters to real statistical parameters of the disease has been developed. The Receiver Operating Characteristic curve has been used to evaluate the quality of the methods as predictive tests.

Keywords: stochastic simulation, predictive classification, pre-symptomatic genetic risk assessment.

Introduction

The purpose of the genetic risk assessment methods is to evaluate the quantitative index which indicates risk of the disease D in the case of an individual's genotype g. The calculation of the risk index in the case of one DNA locus (Single-Nucleotide Polymorphism – SNP) associated with the disease is not complicated and consists of solving of simple system of nonlinear algebraic equations. In this case $g \in \{N, R, R_2\}$ where R denotes risk and non-risk alleles in the diploid, R_2 denotes two risk alleles, N denotes both non-risk alleles. The conditional risk probability $p(D/g)$, the corresponding odds ratio $OR_g = \frac{\mathrm{odds}(p(D/g))}{\mathrm{odds}(p(D/N))}$ [3] or the relative risk $\lambda_g = \frac{p(D/g)}{p(D/N)}$ [7] are risk indexes.

However, chronic non-communicable diseases are caused by a combination of multi-locus genetic risk factors. The risk of multi-locus genotypes have been investigated in many papers (e.g. [3, 7, 2, 4, 8]). The main problem is to evaluate mutual impact of numerous disease-associated loci. If there are k disease-associated loci then there are 3^k different genotypes. Therefore, none of such very expensive statistical evaluations have been conducted until this time. So in all research papers the strong assumption is accepted that impacts of different loci are mutual independent. Also in the papers different assumptions have been accepted about the model of total impact of associated SNPs. The multiplicative model of the overall relative risk has been accepted in [7, 4] and used by Navigenics company. The similar approach was presented in [8]. The product of conditional probabilities was used in [2] and this approach is also equivalent to Navigenics approach. The corporation 23andMe has used another multiplicative model – the product of relative odds [3].

In the genetic databases rather few data about SNPs risk are presented. Therefore the assessment methods' input data should be calculated from known data. Thus, more assumptions should be accepted, e.g. Hardy-Weinberg equilibrium or additive models for log-odds and the impact of risk allele. It is difficult to calculate the error due to such assumptions. Therefore the evaluation of these methods may be done only by the use of experiments.

1 Input of assessment methods

The assessment methods use statistical data about prevalence of genotypes in a specific population and about the risk influenced by one SNP. Such data were estimated by the projects HapMap [1], Wellcome Trust Case Control Consortium (WTCCC) [9] and can be extracted from the knowledge base SNPedia [6]:

- $p(D)$ – an average lifetime risk;
- $OR^i_{R_2}$, OR^i_R – risk odds ratios for homozygous (R_2) and heterozygous (R) genotypes in a i-th locus;
- $f^i_{R_2}$, f^i_R, f^i_N – frequencies of corresponding genotypes in a i-th single locus.

The superscripts for odds ratios OR^i, frequencies f^i, relative risk λ^i and values $\{N^i, R^i, R^i_2\}$ of a genotype identify the locus i.

2 Genetic risk indexes

Different authors have derived different indexes under different assumptions. Suppose, for an individual k diploids SNPs associated with disease are known, therefore his known genotype is a genotype $(g_1 \dots g_k)$, $g_i \in \{N^i, R^i, R^i_2\}$, where g_i is a genotype in i-th locus.

Navigenics Corporation has developed Genetic Composite Index (GCI) [7, 4]:

$$p(D/(g_1 \dots g_k)) \approx \mathrm{GCI} = \frac{p(D)\prod_{i=1}^{k}\lambda^i_{g_i}}{\prod_{i=1}^{k}(f^i_R\lambda^i_R + f^i_{R_2}\lambda^i_{R_2} + f^i_N)}. \tag{1}$$

23andMe Corporation uses a risk index [3] equal to an odds ratio of risk probability in comparison with an average risk:

$$\begin{aligned} OR(g_1 \dots g_k) \equiv \frac{\mathrm{odds}(p(D/g))}{\mathrm{odds}(p(D))} &\approx \prod_{i=1}^{k}\frac{\mathrm{odds}(p(D/g_i))}{\mathrm{odds}(p(D/N^i))} \\ &= \prod_{i=1}^{k}\frac{\lambda^i_{g_i}(1-p(D))}{f^i_R\lambda^i_R + f^i_{R_2}\lambda^i_{R_2} + f^i_N - \lambda^i_{g_i}p(D)}. \end{aligned} \tag{2}$$

The individual is supposed to be at disease developing risk if the risk index exceeds some threshold β. The relative risk $\lambda^i_{g_i}$ values are calculated from data presented in Section 1.

3 Theoretical disease model

We have used the stochastic model [4] of a disease as a sandbox for experiments. The model assumes that the disease is affected by environmental, known genetic and unknown factors that are mutually independent. The risk to develop a disease is simulated by the random variable H:

$$H = \sum_{i=1}^{k} v_i y_i + G + E. \tag{3}$$

Here coefficients $v_i \geq 0$, E is the model of environmental factors and G is the model of undisclosed genetic factors – both are normally distributed random variables with standard deviations σ_e and σ_g, respectively, and a zero means. The random variable y_i ($y_i = 0$, if $g_i = N^i$; $y_i = 1$, if $g_i = R^i$; $y_i = 2$, if $g_i = R_2^i$) is the model of the SNP having large effect and is the Binomially distributed variable $\mathbf{B}(2, p_i)$, where p_i is the frequency of the risk allele $p_i = f_{R_2}^i + 0.5 f_R^i$; the numeral 2 corresponds to 2 trials for an acquiring of a diplotype.

Let us denote the random multidimensional variable by $Y = (y_1, \dots, y_k)$ and its realization by $X = (x_1, \dots, x_k)$. It is assumed that an individual will develop the disease in his lifetime if $H > \alpha$ for an α such that the average lifetime risk $p(D)$ equals the probability $p(H > \alpha)$. The genotype X of an individual is generated according to Binomial distribution $\mathbf{B}(2, p_i)$. Let us denote the set of all generated codes by $\mathbf{X}$.

The parameters of the model (3) should be tuned to real statistical data about the disease. We have no access to clinical data, therefore the model parameters are tuned not to raw data, but to statistics calculated from them. The system of nonlinear equations for tuning model's parameters has been developed and solved.

From the equation (3), independency of risk factors, the formula of total probability, and normality of random variable $G + E$ we have the equation:

$$\begin{aligned} p(D) &= \sum_{\forall X} \left\{ p\left(G + E > \alpha - \sum_i v_i x_i \Big/ Y = (x_i, \dots, x_k) \right) \cdot \prod_{i=1}^{k} p(y_i = x_i) \right\} \\ &= \sum_{x_1=0}^{2} \dots \sum_{x_k=0}^{2} \left\{ 0.5 \left[1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}} \left(\frac{\alpha}{\sigma} - \sum_{i=1}^{k} x_i \frac{v_i}{\sigma} \right) \right) \right] \cdot \prod_{i=1}^{k} B(x_i, 2, p_i) \right\}. \end{aligned} \tag{4}$$

Here $\sigma = \sqrt{\sigma_g^2 + \sigma_e^2}$; $B(x, 2, p) = \frac{2}{x!(2-x)!} p^x (1-p)^{2-x}$ is the probability that the binomial random variable acquires a value x; the ratios α/σ and v_i/σ are assumed to be unknown variables, where $i = 1, 2, \dots, k$.

The k equations follow from the definition of the relative risk

$$\begin{aligned} \lambda_R^j &= \frac{p(D/R^j)}{p(D/N^j)} \\ &= \frac{\sum_{x_1=0}^{2} \dots \sum_{x_{j-1}=0}^{2} \sum_{x_{j+1}=0}^{2} \dots \sum_{x_k=0}^{2} \left\{ 0.5 \left[1 - \operatorname{erf}\left(\frac{\gamma_1^j}{\sqrt{2}} \right) \right] \prod_{i \neq j} B(x_i, 2, p_i) \right\}}{\sum_{x_1=0}^{2} \dots \sum_{x_{j-1}=0}^{2} \sum_{x_{j+1}=0}^{2} \dots \sum_{x_k=0}^{2} \left\{ 0.5 \left[1 - \operatorname{erf}\left(\frac{\gamma_2^j}{\sqrt{2}} \right) \right] \prod_{i \neq j} B(x_i, 2, p_i) \right\}}, \end{aligned} \tag{5}$$

where

$$\gamma_1^j = \left(\frac{\alpha}{\sigma} - \frac{v_j}{\sigma} - \sum_{i \neq j} x_i \frac{v_i}{\sigma}\right), \qquad \gamma_2^j = \left(\frac{\alpha}{\sigma} - \sum_{i \neq j} x_i \frac{v_i}{\sigma}\right), \quad j = 1, 2, \ldots, k.$$

The system of nonlinear equations (4)–(5) can be solved by a numerical method. We have used an approximate solution which gives values of expressions (4)–(5) nearest to statistics $p(D)$ and λ_R^j calculated from real data.

According to the formula (3), the risk to develop a disease for the individual $X \in \mathbf{X}$ is generated by the formula: $H = \sum_i^k v_i x_i + \sigma N(0,1)$, where $N(0,1)$ is a realization of the random variable with the standard normal distribution. The individual is attributed to the positive case sample $\mathbf{X}_d$ (individuals at risk) if $H > \alpha$, that is if $N(0,1) + \sum_i^k x_i \frac{v_i}{\sigma} > \frac{\alpha}{\sigma}$. Otherwise, it is attributed to the negative case sample $\mathbf{X}_n$ (healthy individuals).

4 Experiments

The Receiver Operating Characteristic (ROC) curve is often used to evaluate the quality of a predictive test. The ROC curve is created by plotting the true positive rate $TP(\beta)$ (sensitivity) in the vertical axis against the false positive rate $FP(\beta)$ (fall-out) in the horizontal axis at various threshold β values. The larger is the value of the area under the curve (AUC), the better is the indicator.

The experiments with the model (3) tuned to Type 2 Diabetes (T2D) statistics have been performed. The odds ratios for most significantly associated SNPs for known T2D susceptibility loci for individuals with ancestry from Europe are taken from [10] where results of 3 different scans of populations – Diabetes Genetics Initiative (DGI), Finland-US Investigation (FUSION) and WTCCC are presented. The model (3) has been tuned to the data taken from DGI database.

The system of nonlinear equations (4)–(5) has been expressed as a least squares optimization problem and solved using the Single Agent Stochastic Search (SASS) algorithm [5]. An approximate solution of the system has been used to generate population $\mathbf{X}$ of 100 000 individuals following (3). The generated population has been used to compare reliability of indexes GCIS and OR. The ROC curves of both indexes are presented in Fig. 1(a) as the result of the comparison. One can see from the figure that the difference between the curves is very slight and, therefore, both methods produce very similar result, when the disease model satisfies the same assumptions as risk indexes.

In genetic databases there are no parameters for Lithuanian population; therefore populations with ancestry from Europe are used for Lithuanian patients. In order to investigate the risk indexes' sensibility to the choice of such population the risk indexes and the corresponding ROC curves have been calculated for each of the three European populations, while population $\mathbf{X}$ has been tuned for DGI population. Results of the investigation, presented in Fig. 1(b), show that usage of other population territorially similar to that an individual is taken from, reduces the accuracy of the prediction, though the reduction is slight.

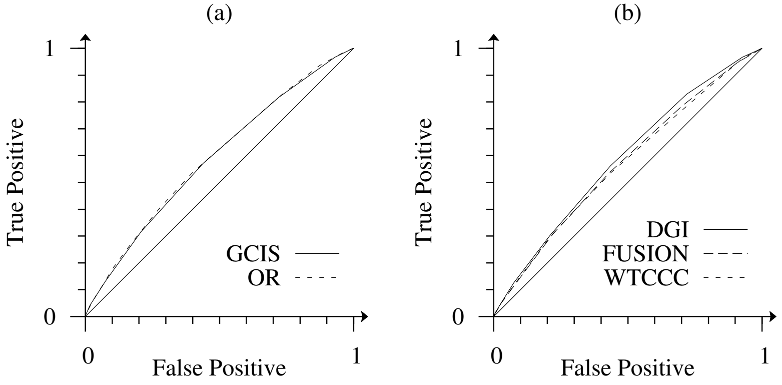

Fig. 1. Comparison of ROC curves: (a) comparison of different indexes; (b) sensibility to the choice of the population.

5 Conclusions

The risk indexes GSIS and OR produce very similar ROC curves in the case when the disease model satisfies the same assumptions as risk indexes. The quality of the risk indexes as a predictive test is low. The indexes are susceptible to the choice of a population territorially similar to that the individual under investigation is taken from. The choice of an improper population reduces the quality of prediction.

References

[1] *HapMap project*. Available from Internet:. http://hapmap.ncbi.nlm.nih.gov/index.html.en. Accessed: 2015-08-21.

[2] Q. Lu and R.C. Elston. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. *Am. J. Hum. Genet.*, **82**(3):641–651, 2008. ISSN 0002-9297. DOI: http://dx.doi.org/10.1016/j.ajhg.2007.12.025.

[3] M. Macpherson, B. Naughton, A. Hsu and J. Mountain. Estimating genotype-specific incidence for one or several loci. *Technical report*, 2007.

[4] B. Padhukasahasram, E. Halperin, J. Wessel, D.J. Thomas, E. Silver, H. Trumbower, M. Cargill and D.A. Stephan. Presymptomatic risk assessment for chronic non-communicable diseases. *PLoS ONE*, **5**(12):e14338, 2010. DOI: 10.1371/journal.pone.0014338.

[5] F.J. Solis and R.J.-B. Wets. Minimization by random search techniques. *Math. Oper. Res.*, **6**(1):19–52, 1981.

[6] *SNPedia*. Available from Internet: http://www.snpedia.com/index.php/SNPedia. Accessed: 2015-02-19.

[7] *The science behind the Navigenics service*. Available from Internet: http://www.navigenics.com/static/pdf/Navigenics%20White%20Paper.pdf. Accessed: 2014-12-16.

[8] N.R. Wray, M.E. Goddard and P.M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, **17**:1520–1528, 2007.

[9] *Wellcome Trust Case Control Consortium (WTCCC).* Available from Internet: http://www.wtccc.org.uk/. Accessed: 2014-12-16

[10] E. Zeggini, L.J. Scott, R. Saxena and B.F. Voigh. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**(5):638–645, 2008.

REZIUMĖ

V. Tiešis, A. Lančinskas, V. Marcinkevičius

Genotipo įtakotos susirgimo rizikos įvertinimo metodų tyrimas

Darbe tiriamos susirgimo rizikos, apspręstos daugelio nukleotidų polimorfizmo ir aplinkos bei elgesio faktorių įvertinimo metodų savybės. Tam panaudotas stochastinis ligos rizikos modelis, sudaryta lygčių sistemą modelio parametrams priderinti prie statistinių ligos parametrų. Palyginti kelių kompanijų naudojami metodai taip pat metodai aprašyti literatūroje. Tirta prognostinio klasifikavimo paklaidų priklausomybė nuo paciento atitikimo populiacijai, kurios statistiniai duomenys naudojami.

Raktiniai žodžiai: stochastinis modeliavimas, prognostinis klasifikavimas, genetinio susirgimo rizikos įvertinimas.