Math. Model. Nat. Phenom.
Vol. 7, No. 3, 2012, pp. 227–252
DOI: 10.1051/mmnp/20127314

Algebraic Methods for Studying Interactions Between Epidemiological Variables

F. Ricceri[1,2], C. Fassino[3], G. Matullo[1,2], M. Roggero[4], M.-L. Torrente[3], P. Vineis[1,5], L. Terracini[4] *

[1] Human Genetics Foundation, Turin, Italy

[2] Department of Genetics, Biology and Biochemistry, University of Turin, Italy

[3] Department of Mathematics, University of Genoa, Italy

[4] Department of Mathematics, University of Turin, Italy

[5] Imperial College, London, UK

Abstract.

Background

Independence models among variables is one of the most relevant topics in epidemiology, particularly in molecular epidemiology for the study of gene-gene and gene-environment interactions. They have been studied using three main kinds of analysis: regression analysis, data mining approaches and Bayesian model selection. Recently, methods of algebraic statistics have been extensively used for applications to biology. In this paper we present a synthetic, but complete description of independence models in algebraic statistics and a new method of analyzing interactions, that is equivalent to the correction by Markov bases of the Fisher's exact test.

Methods

We identified the suitable algebraic independence model for describing the dependence of two genetic variables from the occurrence of cancer and exploited the theory of toric varieties and Gröbner basis for developing an exact independence test based on the Diaconis-Sturmfels algorithm. We implemented it in a Maple routine and we applied it to the study of gene-gene interaction in Gen-Air, an European case-control study. We computed the p-value for each pair of genetic variables interacting with disease status and we compared our results with the standard asymptotic chi-square test.

Results

We found an association among *COMT* Val158Met, *APE1* Asp148Glu and bladder cancer (p-value: 0.009). We also found the interaction among *TP53* Arg72Pro, *GSTP1* Ile105Val and lung cancer (p-value: 0.00035). Leukaemia was observed to significantly interact with the pairs *ERCC2* Lys751Gln and *RAD51* 172 G>T (p-value 0.0072), *ERCC2* Lys751Gln and *LIG4* Thr9Ile (p-value: 0.0095) and *APE1* Asp148Glu and *GSTP1* Ala114Val (p-value: 0.0036).

Conclusion

Taking advantage of results from theoretical and computational algebra, the method we propose was more selective than other methods in detecting new interactions, and nevertheless its results were consistent with previous epidemiological and functional findings. It also helped us in controlling the multiple comparison problem.

In the light of our results, we believe that the epidemiologic study of interactions can benefit of algebraic methods based on properties of toric varieties and Gröbner bases.

Keywords and phrases: polymorphism, interaction, Markov basis, Diaconis-Sturmfels algorithm, independence model, toric variety

Mathematics Subject Classification: 62P10, 62F03, 92B05, 13P10

1. Introduction

Interaction among variables is one of the most relevant topics in epidemiology, particularly in molecular epidemiology. In the last few years, gene-gene (G-G) and gene-environment (G-E) interactions were extensively studied on the assumption of important functional meaning. Most of the recent Genome-Wide Association Studies (GWAS) used a single-locus analysis strategy and the failure to properly address interactions is probably one of the reasons for the failure of GWAS in identifying strong associations in complex diseases. On the contrary, the study of interactions can lead to new possible undescribed direct or indirect functional relationships as reported in [5], where the authors hypothesized that cigarette smoking was associated with an increased risk of colorectal cancer and that this risk might be modified by variants in carcinogen metabolism genes.
A recent review [6] describes the methods that are used for testing interactions. There are three main kinds of analysis: regression analysis, data mining approaches and Bayesian model selection.
The traditional parametric method to analyze interactions is logistic regression. Using this method, it is possible to estimate the main effects and the interaction terms and to test if interaction terms contribute or not to the model. This method is affected by a series of problems, especially in analyzing genetic databases with a large number of variables compared to the number of subjects.
Another method that was developed for the analysis of G-G and G-E interactions is Multifactor Dimensionality Reduction (MDR) [22, 43], that is a nonparametric, model-free algorithm that reduces the dimensionality of multilocus information, to improve the identification of polymorphism combinations associated with disease risk. This method has been extensively used both in case-control studies [27] and in discordant-sib-pair studies [30]. The main limitation of this approach is that, in the case of extensive analyses of more than few hundreds of variables, it is too time consuming. The other problem is that in many instances, results obtained using MDR were different from results obtained using other methods [27, 35].
Recently, Bayesian methods for testing interactions have been presented, in particular the Bayesian Epistasis Association Mapping (BEAM) [52] and profile regression analysis [31].
Algebraic applications to statistics have been extensively described by different authors [10,34,38,39] and different papers on their applications to biology have been published [4, 23, 33, 36, 48]. An application to epistasis is provided by [26].
In this paper we describe a new method to analyze epidemiological interactions, using algebraic statistics and computational algebra. This method has been used previously to test the association between two variables [39] and, in this case, it is equivalent to the correction by Markov bases of Fisher's exact test. We propose an extension to three-way biological interactions.
The main advantages of algebraic methods compared to classical methods are their parsimony (lower number of degrees-of-freedom) and their being assumption-free, i.e. they do not specify the mathematical form of the interaction. Furthermore, compared to other assumption-free methods such as MDR, they are not empirical but are based on a sound theoretical background.

*Corresponding author. E-mail: lea.terracini@unito.it

2. Material and methods

2.1. The mathematical model

We were interested in testing the effect of the combination of two polymorphims on the risk of cancer. Then the most suitable model to apply is the model of independence of two random variables, say X_1, X_2, from a third one X_3 which is a binary variable denoting the presence or absence of cancer; in symbols

$$P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i, X_2 = j)P(X_3 = k) \tag{2.1}$$

This model is exactly the standard log-linear model (it is sufficient to take logarithms in formula (2.1)); we shall write it in multiplicative form in order to emphasize its toric structure.
We refer to the Appendix for a general description of toric models, their corresponding formulation as log-linear models, and their algebraic and geometrical properties.
The parametric equations of model (2.1) are of the form

$$p_{ijk} = \theta_{ij}\mu_k.$$

Depending on the cases, X_1 and X_2 can have values in $\{0, 1\}$ or in $\{0, 1, 2\}$. We refer to the case where both X_1 and X_2 are ternary variables as case TT, to the case where one of them is binary as case TB, and to the case where both are binary as case BB. We associate to this model the ideal $I_{12,3}$ (see Appendix, Section 5.2) that contains all polynomials in the variables p_{ijk} that identically vanish on the points given by the parametrization. It is a toric ideal and a Gröbner basis for it is given by binomials of the form

$$p_{i_1j_1k_1}p_{i_2j_2k_2} - p_{i_1j_1k_2}p_{i_2j_2k_1}. \tag{2.2}$$

The Gröbner basis consists of 36 binomials in case TT, of 15 binomials in case TB, and of 6 binomials in case BB.

As explained in the Appendix the likelihood is easily maximized for every toric model, using algebraic and/or numerical techniques.

We used the Pearson coefficient

$$\sum \frac{(\text{obs} - \exp)^2}{\exp}$$

as our test statistic.

We tested interaction of each pair of polymorphisms on cancer using a method, originally proposed by Diaconis and Sturmfels [9], which belongs into the category of exact methods. It is known (see [1, 2]) that exact methods are useful in situations where the asymptotic assumptions are not met, and so the asymptotic p-value is not a good approximation for the true p-value: for example when the size of samples is small, or when the data distribution is sparse. In their more general form, they consist in comparing the test statistic of the observed sample with that of every element in a reference finite set $\mathcal{Y}$ consisting of all possible samples having the same sufficient statistics (the MLE in our case). If $\mathcal{Y}_0$ denotes the subset of $\mathcal{Y}$ consisting of samples having the test statistic greater than or equal to the observed one, the exact p-value is defined as the sum of probabilities of elements in $\mathcal{Y}_0$. In our case, probability is uniformly distributed on the reference set $\mathcal{Y}$ (and thus also on $\mathcal{Y}_0$), so that the p-value is simply the ratio of cardinalities

$$\frac{|\mathcal{Y}_0|}{|\mathcal{Y}|}.$$

As noted by Diaconis and Sturmfels [9] (see also the discussion in [26]) the reference set is in general very large, and then difficult to enumerate even for very small data sets, where the asymptotic approximation is heavily inadequate. Then Monte Carlo methods can be used in order to sample from it according to a fixed distribution. Using the Gröbner basis of the ideal associated to the model, Diaconis-Sturmfels algorithm allows us to obtain in a simpler way a Monte Carlo sampling (see Appendix, Section 5.5).

We applied D-S algorithm to each pair of SNPs (Single Nucleotide Polymorphisms) in the database, and computed the corresponding Monte Carlo p-value. The number of pairs is quadratic in the number of variables, so this is computationally expensive; however the Gröbner basis only depends on the model and on the number of values of X_1, X_2, so it is computed only once for each of the three cases TT, TB, BB.

We implemented the Diaconis-Sturmfels test in a Maple routine. The software is available at our website:
http://www.unito.it/unitoWAR/page/dipartimenti1/D005/D005_merima1.

2.2. Study population

Our method has been tested in the Gen-Air study [49].
Gen-Air is a case-control study nested in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, that is a multicenter European study, in which more than 500 000 healthy volunteers were recruited in 10 European countries (France, Denmark, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and UK) corresponding to 23 recruitment centers [41] .
The aim of Gen-Air is to study the relationship between some types of cancer and air pollution or environmental tobacco smoke (ETS); for this reason only non-smokers or former smokers since at least 10 years have been included. Cases are subjects with bladder, lung, oral, pharyngeal or laryngeal cancers and leukemia, diagnosed after recruitment. For each case, three controls were matched for exposure assessment and the analysis of questionnaire data, and two controls for laboratory analyses. Matching criteria were gender, age (± 5 years), smoking status (never or former smoker), country of recruitment, and follow-up time.
In the present study, we analyzed 124 cases of bladder cancer, 116 of lung cancer, 169 of leukemia and 757 controls (breaking the matching between cases and controls) with available blood samples and successful DNA extraction and genotype analysis. In the analyses we included 35 polymorphisms in 28 genes (Table 1).
Motivations for the choice of relevant polymorphisms and methods of genotyping are described elsewhere [27, 29, 37].

3. Results

We observed that 10.000 iterations of the Diaconis-Sturmfels algorithm were able to discriminate between significant (p-value < 0.05) and non-significant results. We also observed that 100.000 iterations were able to give consistent results at the fourth decimal figure. For this reason we tested all the triplets using 10.000 iterations and we repeated the analysis using 100.000 iterations for the triplets that showed a p-value less than 0.10.
In Tables 2, 3 and 4 we show the results of G-G interactions for bladder cancer, lung cancer and leukaemia, respectively. In each table, results in the diagonal represent the association between each SNP and the corresponding cancer. It is obvious that significant results in single association lead to significant results for almost all the interactions involving that particular SNP, so we did not consider them because they are not proper interactions. Significant Monte Carlo p-values are emphasized in bold in the tables.
For bladder cancer the main result was the association between *COMT* Val158Met, *APE1* Asp148Glu and the case-control status (p-value: 0.009). For lung cancer, the most significant result concerned the interaction among *TP53* Arg72Pro, *GSTP1* Ile105Val and the disease (p-value: 0.00035). For leukaemia we observed three strongly significant results: the interaction between leukaemia and the pairs *ERCC2* Lys751Gln and *RAD51* 172 G>T (p-value 0.0072), *ERCC2* Lys751Gln and *LIG4* Thr9Ile (p-value: 0.0095) and *APE1* Asp148Glu and *GSTP1* Ala114Val (p-value: 0.0036).
In Table 5 we show the results which are significant either for the chi-square goodness-of-fit test or for our test, or both. In about 30% of analyses, the chi-square failed to produce results because of zero values in some categories. In all other cases, we can see that the associations detected with our method remain

significant using the chi-square test, and that our method results to be more conservative. In many cases, chi-square was not the recommended test because in at least one category the expected value was lower than 5.

4. Discussion

In this paper we show a new application of algebraic statistical methods to the study of interactions in epidemiology. The interpretation of independence models as toric models and the corresponding algebraic and geometrical theory allowed us to construct an algorithm for testing gene-gene interaction in epidemiology.

The main advantages of our approach compared to other existing approaches are its parsimony, its being assumption-free and its simplicity. Furthermore, although it is based on a rigorous theoretical background, developed by world famous mathematicians, the underlying idea is extremely simple, natural and easy to implement.

Using this approach we were able to identify five G-G interactions for cancer risk.

For lung cancer, we detected an interaction between the two polymorphisms *TP53* Arg72Pro and *GSTP1* Ile105Val. This result is supported by previous findings: Gorij *et al.* found that the Val/Ile and Val/Val forms in this SNP in the *GSTP1* gene were associated with lack of TP53 expression in Basal Cell Carcinoma [21]. Furthermore, it has been demonstrated that *GSTP1* gene is a heretofore unrecognized downstream transcriptional target of the tumor suppressor p53 and it is transcriptionally activated by p53 through binding of p53 to a p53-binding motif present in the *GSTP1* gene [25]. This interaction was also found by Manuguerra *et al.* analyzing GenAir data using MDR [27].

Similarly, the interaction between *COMT* Val158Met and *APE1* Asp148Glu for the risk of bladder cancer was found in MDR analysis of four variables. COMT is a well-known protein that catalyses the methylation of various endobiotic and xenobiotic substances and prevents quinone formation and redox cycling. The unfavourable COMT variant might thus lead to an increased oxidative DNA damage that can induce base excision repair (BER) pathway, interacting in particular with *APE1* activity [3]. Moreover, in an independent study [28] p53 mutations were observed to occur in subjects with the *COMT*.

Three of the five genes that we found to interact in leukaemia (*GSTP1, RAD51* and *APE1*), were also detected by the MDR analysis. Regarding the association between *ERCC2* Lys751Gln and *RAD51* 172 G>T, a recent analysis on the crystal structure of *ERCC2* [15] showed that its catalytic core was composed by four domains, one of which resembled the ATPase domain in *RAD51*. The interaction between *ERCC2* Lys751Gln and *LIG4* Thr9Ile is not surprising. In fact it is known that genes involved in the NER DNA-repair pathway interact with genes in DSBR repair pathway [53], particularly in the interstrand crosslink repair [51].

Our method belongs to the family of permutation tests, that grew out of the works of Fisher [18] and that is an alternative to Bonferroni's correction for multiple comparisons. The general idea is that reference distributions for the test statistics for the observed data are derived by rearranging the original data points and the calculation of all possible values of that test statistics. The assumption is that "if all tests are equivalent then the test statistic should be the same even if the data points are jumbled"[42]. The specific purpose of these methods is to compute how often a given p-value would be found by chance if the study were repeated without any associations. The properties of permutations tests (in relation to the reduction of I type error) were widely described elsewhere ([8, 13, 14, 20, 32, 44]). Moreover, in the contest of Genome-Wide association, many new permutation tests were recently proposed ([11, 12, 24]) and a very recent paper by Wang *et al.* describes how these tests can be used in gene-environment interaction studies ([50]). For this reason we can suggest that our method helps in disentangling the problem of multiple comparison in a robust way, regardless the underlying distribution.

As a supplementary confirm of this, we compared our results with the asymptotic chi-square goodness-of-fit test. In most of the cases, the chi-square approach failed to produce reliable results (because of sparseness of data). When chi-square was applicable, it seemed that our technique was more conservative, suggesting a better control for multiple comparisons.

The main problem of our approach is that implementation required running the algorithm for each pair with a large number of iterations. For example, in the context of the study of interactions in a GWAS, this method at the moment would be computationally too expensive. Another problem is that interactions are only tested but not quantified (for example in terms of Odds Ratio). It would be interesting to develop a measure of interaction by using algebraic techniques; some suggestions arise from the recent papers by Fassino and Torrente [16]. Another limit of our approach is that we confined our analysis to the study of three-way interactions. Sturmfels' hierarchical models described in the Appendix could be applied in order to generalize our method to the interaction of four or more variables.
In conclusion, interactions among genetic variants were successfully detected using our method that also helped in tackling the multiple comparison problem. Furthermore, the different kinds of interaction are very well outlined in the algebraic context and Markov bases are easily computed by Buchberger algorithm.
In the light of our results, we believe that the epidemiologic study of interactions can benefit of algebraic methods based on properties of toric varieties and Gröbner bases.

5. Appendix: mathematical background

5.1. Toric statistical models

Toric models (or log-linear models) are a very important class of statistical models. In the following we summarize the part of the theory which is relevant for our work; for details see [36, 45].

Let $A = (a_{ij})$ be a $d \times n$ matrix having non negative integral entries, and such that the sum of each column is the same:

$$\sum_{i=1}^{d} a_{i1} = \sum_{i=1}^{d} a_{i2} = ... = \sum_{i=1}^{d} a_{in} \tag{5.1}$$

The j-th vector column $\mathbf{a}_j = (a_{1j}, ..., a_{dj})$ of A represents the monomial

$$\theta^{\mathbf{a}_j} = \theta_1^{a_{1j}} \theta_2^{a_{2j}} ... \theta_d^{a_{dj}}.$$

The toric model M_A associated to A is given by the parametrization

$$\begin{aligned} \varphi : \theta = \mathbf{R}^d &\longrightarrow \mathbf{R}^n \\ \theta &\longmapsto (\theta^{\mathbf{a}_1}, \theta^{\mathbf{a}_2}, ..., \theta^{\mathbf{a}_n}). \end{aligned}$$

that is it is defined by the parametric equations

$$\begin{cases} p_1 = \theta^{\mathbf{a}_1} = \theta_1^{a_{11}} \theta_2^{a_{21}} ... \theta_d^{a_{d1}} \\ \quad \vdots \\ p_n = \theta^{\mathbf{a}_n} = \theta_1^{a_{1n}} \theta_2^{a_{2n}} ... \theta_d^{a_{dn}} \end{cases} \qquad \theta_1, ..., \theta_d > 0 \tag{5.2}$$

The hypothesis (5.1) ensures that all monomials $\theta^{\mathbf{a}_i}$ have the same degree.

By the elimination theorem (see for example [7, p. 113]), we can eliminate the parameters θ_i and we obtain an ideal I_A in the polynomial ring $\mathbb{Q}[p_1, ..., p_n]$ which is called the ideal of the toric model. In other words, we get algebraic equations for the variety V_A that is the minimal algebraic variety containing the points given by the parametrization (called the Zariski closure of this set of points). In general the Zariski closure may contain more points than the original parametrized analytical variety. It is well-known that I_A has a special set of generators (a Gröbner basis) consisting of homogeneous binomials [45]. The points $(p_1, ..., p_n)$ in V_A such that $p_i \geq 0$ for $i = 1, ..., n$ and $\sum_{i=1}^{n} p_i = 1$ are the statistically relevant points of the variety; each of such point may be regarded as a particular probability distribution in the model.

5.2. Independence models

An important class of toric models is provided by independence models of a finite set of random variables, that we briefly describe. Some of them has been studied in [26].

Independence models of two random variables

The most simple example is the independence model of two random variables. For each natural number d, we denote by $[d]$ the set $\{1, ..., d\}$. Let d_1, d_2 be two natural numbers. We introduce variables $p_{ij}, i \in [d_1], j \in [d_2]$, $\theta_1, ..., \theta_{d_1}, \mu_1, ..., \mu_{d_2}$ and consider the model defined by the parametric equations

$$\left\{ p_{ij} = \theta_i \mu_j \qquad (i,j) \in [d_1] \times [d_2], \theta_i, \mu_j \in \mathbf{R}_{>0} \right. \tag{5.3}$$

By taking logarithms in the parametric equations, we get the standard log-linear model:

$$\log(p_{ij}) = \lambda_i^{X_1} + \lambda_j^{X_2}.$$

For each value of the vector of parameters $\theta = (\theta_1, ..., \theta_{d_1}, \mu_1, ..., \mu_{d_2})$ we obtain a point with positive coordinates $(p_{ij}(\theta), i \in [d_1], j \in [d_2])$ on the corresponding variety. If we further impose that $\sum_{i,j} p_{ij}(\theta) = 1$, then we can regard the function

$$p(i,j) = p_{ij}(\theta)$$

as the joint density function of two independent random variables X_1, X_2 with values in $[d_1]$ and $[d_2]$ respectively. The density functions of X_1, X_2 will be respectively

$$p(i) = \theta_i \sum_j \mu_j \qquad i = 1, ..., d_1$$

$$q(j) = \mu_j \sum_i \theta_i \qquad j = 1, ..., d_2$$

Conversely, if X_1, X_2 are two independent random variables in $[d_1]$ and $[d_2]$ respectively, then their joint density function $(P(X_1 = i, X_2 = j))$ is represented by a point in the model.

Then the model defined by equations (5.3) represents the family of all joint random variables of two independent random variables X_1 and X_2 with values in $[d_1]$ and $[d_2]$ respectively.

By eliminating the parameters $\theta_1, ..., \theta_{d_1}, \mu_1, ..., \mu_{d_2}$ we get the ideal

$$I := (p_{i_1 j_1} p_{i_2 j_2} - p_{i_1 j_2} p_{i_2 j_1} \mid 1 \le i_1 < i_2 \le d_1, 1 \le j_1 < j_2 \le d_2)$$

The corresponding variety is the well-known **Segre variety**. It is easily seen from the generators of I that the Segre variety consists of those matrices $\{p_{ij}\}$ having all minors of order 2 equal to 0, thus having rank 1.

Independence models for three random variables

Consider now three discrete (categorical) random variables X_1, X_2, X_3 with values in finite sets $[d_1]$, $[d_2]$ and $[d_3]$ respectively. Then we can construct the joint random variables $X_{12} = (X_1, X_2)$, $X_{13} = (X_1, X_3)$, $X_{23} = (X_2, X_3)$ and $X = (X_1, X_2, X_3)$. Denote by P the probability function. Besides the problem of studying the complete independence of the three original variables X_1, X_2, X_3, one might be interested in the study of independence of some subset of variables in $\{X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X\}$. For some subset this problem does not make sense: for example it is obvious that X_1 cannot be independent from X_{12}, and each variable is dependent from X. In [47] Sturmfels analyzes the five significant types of independence of three random variables:

– **Complete independence:** The complete independence model is the classical independence of the three variables, which must be subjected to the condition

$$P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i)P(X_2 = j)P(X_3 = k)$$

for $(i, j, k) \in [d_1] \times [d_2] \times [d_3]$.
In order to give parametric equations of the model, we introduce variables p_{ijk} and parameters $\theta_1, ..., \theta_{d_1}$, $\mu_1, ..., \mu_{d_2}$, $\nu_1, ..., \nu_{d_3}$ in the space $\mathbf{R}^d_{>0}$ and consider the $d_1d_2d_3$ equations of the form

$$\{p_{ijk} = \theta_i\mu_j\nu_k$$

The ideal of the corresponding variety is denoted by $I_{1,2,3}$.
By taking logarithms in the parametric equations, we get the standard log-linear model:

$$\log(p_{ijk}) = \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^{X_3}.$$

- **Independence of** (X_1, X_2) **from** X_3**:** It is the model that we will adopt in our study. It is defined by the condition

$$P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i, X_2 = j)P(X_3 = k)$$

for$(i, j, k) \in [d_1] \times [d_2] \times [d_3]$.
The model involves $d_1d_2+d_3$ parameters $\theta_{ij}, (i, j) \in [d_1]\times[d_2]$, $\mu_1, ..., \mu_{d_3}$ and has parametric equations

$$\{p_{ijk} = \theta_{ij}\mu_k; \tag{5.4}$$

the ideal of the corresponding variety is denoted $I_{12,3}$.
It is shown in [47] that a Gröbner basis for $I_{12,3}$ is provided by the binomials of the form

$$p_{i_1j_1k_1}p_{i_2j_2k_2} - p_{i_1j_1k_2}p_{i_2j_2k_1}. \tag{5.5}$$

In logarithmic notation, the model is given by

$$\log(p_{ijk}) = \lambda_{ij}^{X_1X_2} + \lambda_k^{X_3}.$$

In our specific cases (TT, TB and BB), we imposed on the parameters the order

$$\theta_{00}, \theta_{01}, \theta_{02}, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{20}, \theta_{21}, \theta_{22}, \mu_0, \mu_1,$$

and on the variables p_{ijk} the lexicographic order. Then the matrix A associated to our models is an 11×18 matrix in case TT (8×12 and 6×8 respectively in cases TB and BB) of the form

$$A = \begin{pmatrix} 1\ 1\ 0\ 0\ ...\ 0\ 0 \\ 0\ 0\ 1\ 1\ ...\ 0\ 0 \\ \vdots \\ 0\ 0\ 0\ 0\ ...\ 1\ 1 \\ 1\ 0\ 1\ 0\ ...\ 1\ 0 \\ 0\ 1\ 0\ 1\ ...\ 0\ 1 \end{pmatrix};$$

we see that the rank of A is one less the number of rows of A in each case.

- **Independence of** X_1, X_2 **conditionally to** X_3
We say that X_1 and X_2 are **independent conditionally to** X_3 (in symbols $X_1 \coprod X_2|X_3$) if

$$P(X_1 = i, X_2 = j \mid X_3 = k) = P(X_1 = i \mid X_3 = k)P(X_2 = j \mid X_3 = k)$$

for every i, j, k.
The parametric equations of the model are

$$\{p_{ijk} = \theta_{ik}\mu_{jk}.$$

The ideal of the corresponding variety is denoted by $I_{12,23}$.
In logarithmic notation, the model is given by

$$\log(p_{ijk}) = \lambda_{ik}^{X_1X_3} + \lambda_{ij}^{X_2X_3}.$$

- **No three way interaction**
  It is the model that parametrizes those distributions where all possible interactions involve only two of the three variables. The parametric equations of the model are of the form

$$\{p_{ijk} = \theta_{ij}\mu_{ik}\nu_{jk}.$$

  In logarithmic notation:

$$\log(p_{ijk}) = \lambda_{ij}^{X_1X_2} + \lambda_{ik}^{X_1X_3} + \lambda_{jk}^{X_2X_3}.$$

  The ideal of the corresponding variety is denoted by $I_{12,13,23}$.
- **Models not depending on one or more variables**
  In the previous models the value p_{ijk} depends on all indices i, j, k. Consider now the following models

$$\{p_{ijk} = \theta_i \quad \text{or} \ \log(p_{ijk}) = \lambda_i^{X_1} \tag{5.6}$$

$$\{p_{ijk} = \theta_{ij} \quad \text{or} \ \log(p_{ijk}) = \lambda_{ij}^{X_1X_2} \tag{5.7}$$

$$\{p_{ijk} = \theta_i\mu_j \quad \text{or} \ \log(p_{ijk}) = \lambda_i^{X_1} + \lambda_j^{X_2} \tag{5.8}$$

  We see that the joint probability does not depend on the value of X_3 (and in the first case not even on the value of X_2). Therefore X_3 (and also X_2 in the first case) are uniformly distributed and independent from the remaining variables. The second case allows dependence between X_1 and X_2, while in the third case X_1 e X_2 are independent. The ideals of the corresponding variables are denoted by $I_1, I_{12}, I_{1,2}$ respectively.

Each model gives rise to a toric ideal in the ring $Q[\{p_{i,j,k}\}]$, with $(i, j, k) \in [d_1] \times [d_2] \times [d_3]$. A Gröbner basis $\mathcal{B}$ for the ideal, w.r.t. a suitable monomial order, can be computed by Buchberger algorithm [7].

Independence of four or more random variables: hierarchical models
A generalization of the previous analysis to the case of more than three random variables is described in [36]. Suppose to have random variables $X_1, ..., X_n$ with values in $[d_1], ..., [d_n]$ respectively. Then we can construct different independence models for these variables, as follows.
Let Σ be a collection of subsets of $[n] = \{1, .., n\}$ such that $\sigma_1 \not\subseteq \sigma_2$, for every $\sigma_1, \sigma_2 \in \Sigma$. For every $\sigma = \{r_1, ..., r_k\} \in \Sigma$ we introduce a set of parameters $\theta_{\sigma,i_1,...,i_k}$, with $(i_1, ..., i_k)$ varying in $[d_{r_1}] \times ... \times [d_{r_k}]$ and consider the model given by the parametric equations:

$$\mathcal{M}_\Sigma : \quad \left\{ p_{i_1,...,i_n} = \prod_{\sigma\in\Sigma,\ \sigma=\{r_1,...,r_k\}} \theta_{\sigma,i_{r_1},...,i_{r_k}} \right.$$

The corresponding ideal in $\mathbf{Q}[p_{i_1,...,i_n} \mid (i_1, ..., i_n) \in [d_1] \times ... \times [d_n]]$ is denoted by I_Σ.

Example 5.1. In the case $n = 4$, put $\Sigma = \{\{1, 3\}, \{2, 3\}, \{4\}\}$,$A = \{1, 3\}$, $B = \{2, 3\}$, $C = \{4\}$; then the corresponding model has parametric equations

$$\mathcal{M}_\Sigma : \quad \{p_{i_1,i_2,i_3,i_4} = \theta_{A,i_1,i_3}\theta_{B,i_2,i_3}\theta_{C,i_4}$$

5.3. Lattices, fibers and Markov basis

Following [45] we shall call a **lattice** a subgroup of $\mathbf{Z}^n$.

Given a lattice $\mathcal{L}$ and a vector $\mathbf{u}$, the **fiber** of $\mathbf{u}$ is the set

$$\begin{aligned}\mathcal{F}(\mathbf{u}) &= (\mathbf{u} + \mathcal{L}) \cap \mathbf{N}^n \\ &= \{\mathcal{L} \in \mathbf{N}^n \mid \mathbf{u} - \mathcal{L} \in \mathcal{L}\}\end{aligned}$$

It is easy to see that if the only non-negative vector (that is such that every component is non-negative) in $\mathcal{L}$ is the zero vector then every fiber is a finite set. Every subset $\mathcal{B}$ of $\mathcal{L}$ gives each fiber $\mathcal{F}$ a natural structure $\mathcal{F}_\mathcal{B}$ of undirected graph: namely, there is an edge between two vectors $\mathbf{u}$ and $\mathbf{v}$ if $\mathbf{v} = \pm\mathbf{b} + \mathbf{u}$ for some $\mathbf{b} \in \mathcal{B}$. A finite subset $\mathcal{B}$ of $\mathcal{L}$ is said to be a **Markov basis** of $\mathcal{L}$ if the graphs $\mathcal{F}(\mathbf{u})_\mathcal{B}$ are connected for every $\mathbf{u} \in \mathbb{N}^n$.

Now let A be a $d \times n$ matrix defining a toric model. We can regard A as a linear application $\mathbf{Z}^n \longrightarrow \mathbf{Z}^d$; then its kernel is a lattice in $\mathbf{Z}^n$, that we denote $\mathcal{L}_A$. Observe that condition (5.1) implies that the only non-negative vector in $\mathcal{L}_A$ is the zero vector: indeed every vector in $\mathcal{L}_A$ must be orthogonal to the vector $(1, ..., 1)$. Therefore for every $\mathbf{u} \in \mathbb{N}^n$ the fiber $\mathcal{F}_A(\mathbf{u})$ is a finite set. Let $\mathcal{G} = \{g_1, ..., g_r\}$ be a reduced Gröbner basis of I_A; then each g_i is a binomial, so we can write

$$g_i = p_1^{u_{i1}} ... p_n^{u_{in}} - p_1^{v_{i1}} ... p_n^{v_{in}} \qquad \text{with } u_{ij}, v_{ij} \in \mathbb{N}$$

and since the ideal I_A is prime we may assume that the two vectors $\mathbf{u_i} = (u_{i1}, ..., u_{in})$ and $\mathbf{v_i} = (v_{i1}, ..., v_{in})$ have disjoint supports. The following result, proved in [9], is crucial:

Theorem 5.2. *If $\mathcal{G}$ is a reduced Gröbner basis of I_A then $\mathcal{B} = \{\mathbf{u_i} - \mathbf{v_i} \mid i = 1, ..., r\}$ is a Markov basis of $\mathcal{L}_A$.*

By Theorem 5.2, Markov bases can be effectively computed by Buchberger algorithm and its improvements (see [7, Chap. 2, Section 7]; see also the website related to [26], that provides Markov bases for the models they use).

A very useful application of the above result is the Diaconis-Sturmfels algorithm which provides a method for sampling from a fiber according to a given distribution. It is a variant of Metropolis-Hastings algorithm which uses a Markov basis $\mathcal{B}$ to generate moves on the fiber. For details see [9].

Diaconis-Sturmfels algorithm
Let $\mathcal{B}$ be a Markov basis for $\mathcal{L}$ and let σ be a probability distribution on a fiber $\mathcal{F}$.

Step 1. Start from $\mathbf{h}_0 = \mathbf{k}_0 \in \mathcal{F}$;
Step 2. Randomly choose $\mathbf{b} \in \mathcal{B}$ and $\epsilon \in \{\pm 1\}$.
Step 3. If $\mathbf{h}_r + \epsilon\mathbf{b} \notin \mathbf{N}^n$ then set $\mathbf{h}_{r+1} = \mathbf{h}_r$; else, move to $\mathbf{h}_{r+1} = \mathbf{h}_r + \epsilon\mathbf{b}$ with probability $\min\{\frac{\sigma(\mathbf{h}_r+\epsilon\mathbf{b}}{\sigma(\mathbf{h}_r)}), 1\}$, i.e. put

$$\mathbf{h}_{r+1} = \begin{cases} \mathbf{h}_r + \epsilon\mathbf{b} & \text{with probability } \mu = \min\{\frac{\sigma(\mathbf{h}_r+\epsilon\mathbf{b}}{\sigma(\mathbf{h}_r)}), 1\} \\ \mathbf{h}_r & \text{with probability } 1 - \mu \end{cases}$$

Since the graph $\mathcal{F}_B$ is connected, the underlying Markov process is connected, reversible and aperiodic [9, Lemma 2.1]; therefore by the Ergodic theorem for Markov Chains, σ is the unique stationary distribution, and uniformly for $\mathbf{h} \in \mathcal{F}$,

$$\lim_{n\to\infty} P(\mathbf{h}_n = \mathbf{h}) = \sigma(\mathbf{h}).$$

5.4. Maximum likelihood estimation for toric models

For toric models the maximum-likelihood estimation (MLE) of a frequency vector $\mathbf{k}_0$ of a set of observations can be determined in an algebraic way.

Given a set of observations let $\mathbf{k} = (k_1, ..., k_n)$ be the associated frequency vector. For a toric model M_A of the form (5.2) the log-likelihood function is linear in the logarithm of the parameters:

$$\ell_{(k_1,...,k_n)}(\theta_1, ..., \theta_d) = \sum_{i=1}^{d} \sum_{j=1}^{n} k_j a_{ij} \log(\theta_i)$$

The log-MLE is thus obtained by maximizing the function ℓ with the constraint

$$\theta_1^{a_{11}} ... \theta_d^{a_{d1}} + ... + \theta_1^{a_{1n}} ... \theta_d^{a_{dn}} = 1.$$

This can be done using the method of Lagrange multipliers. It is shown in [36] that the critical points of the lagrangian function associated to the optimization problem above is the set of solutions of the system

$$\begin{cases} a_{11}\theta_1^{a_{11}}...\theta_d^{a_{d1}} + a_{12}\theta_1^{a_{12}}...\theta_d^{a_{d2}} + ... + a_{1n}\theta_1^{a_{1n}}...\theta_d^{a_{dn}} = \frac{1}{k}\sum_{j=1}^n k_j a_{1j} \\ \vdots \\ a_{d1}\theta_1^{a_{11}}...\theta_d^{a_{d1}} + a_{d2}\theta_1^{a_{12}}...\theta_d^{a_{d2}} + ... + a_{dn}\theta_1^{a_{1n}}...\theta_d^{a_{dn}} = \frac{1}{k}\sum_{j=1}^n k_j a_{dj} \end{cases}$$

where $k = \sum_{i=1}^n k_i$. By using the parametrization (5.2) we can regard this set as the set of points $\mathbf{p} = (p_1, ..., p_n)$ in the zero-set of the ideal I_A which are solutions of the linear system

$$A\mathbf{p} = \frac{1}{k}A\mathbf{k}, \tag{5.9}$$

that is the set of points in the variety of the ideal $J_A = I_A + (A\mathbf{p} - \frac{1}{k}A\mathbf{k})$. The solutions which are relevant for the statistical problem are those having $p_i \geq 0$ for $i = 1, .., n$; a theorem of Birch about convex polytopes [36, Theorem 1.10] ensures that there is a unique such solution. This is an easy fact if the frequency vector has strictly positive entries; when some k_i is zero, it is a consequence of the properties of the moment map, see [19, §4.2] and [17, Theorem 4 and Appendix A]. Therefore, the MLE for a toric model is well-defined for every set of observations; moreover, it can be easily computed from the ideal J_A using algebraic methods [46]. In the case of our independence model it can be proved that this point also belongs to the image of the parametrization, if we allow parameters vary in $[0, +\infty)$.

In general there is not a closed formula expressing the MLE as a function of the matrix A and the vector $\mathbf{k}$ of observed frequencies. However, case by case, the solution can be computed (this is possible using the most common symbolic calculation softwares) with the algebraic and/or numerical techniques.

Equation (5.9) has a very important consequence: suppose to have two set of observations $\mathcal{U}_1$ and $\mathcal{U}_2$ of the same size k and let $\mathbf{k_1}$ and $\mathbf{k_2}$ be their respective relative frequency vectors. Then $\mathcal{U}_1$ and $\mathcal{U}_2$ have the same MLE if and only if $\mathbf{k_1}$ and $\mathbf{k_2}$ are in the same fiber $\mathcal{F}$ of the lattice $\mathcal{L}_A$. If this is the case and $\mathcal{B}$ is a Markov basis of $\mathcal{L}_A$, then there will be a finite sequence of elements $\mathbf{u}_0, \mathbf{u}_1, ...\mathbf{u}_s$ such that $\mathbf{u}_0 = \mathbf{k_1}, \mathbf{u_s} = \mathbf{k_2}$ and for $i = 1, ..., s$, $\mathbf{u_i} = \mathbf{u_{i-1}} \pm \mathbf{b}_i$ for some $\mathbf{b}_i \in \mathcal{B}$. This fact gives the theoretical foundation for the exact inference method described below.

5.5. Exact methods for goodness-of-fit tests and the Diaconis-Sturmfels algorithm

Let A be a matrix $d \times n$, and M_A be the toric model defined by equations (5.2). As explained above, the model describes a family of discrete distributions on the set $\{1, .., n\}$ and thus observations can belong to n different classes. In our application n is $3 \times 3 \times 2$ in case TT, $3 \times 2 \times 2$ in case TB and $2 \times 2 \times 2$ in case BB.

Suppose to have a set $\mathcal{U}_0$ of k observations, let $\mathbf{k}_0 = (k_1, ..., k_n)$ be the observed frequency vector and compute the MLE $\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_n)$.

Consider the Pearson coefficient:

$$C_{\mathcal{U}_0} = C_{\mathbf{k}_0} = \sum_{i=1}^n \frac{(k\hat{p}_i - k_i)^2}{k\hat{p}_i}.$$

Notice that by the proof of Birch theorem in [36, Theorem 1.10] $\hat{p}_i = 0$ implies $k_i = 0$ so that we can assume $\frac{(k\hat{p}_i - k_i)^2}{k\hat{p}_i} = 0$ in this case.

The standard asymptotic χ-square test is based on the fact that C, regarded as a random variable, tends in law to the limit distribution $\chi^2(n - 1 - d')$ where d' is the number of parameters estimated from the sample. In the toric case formula (5.9) shows that $d' = \text{rank}(A) - 1$. Then one rejects the goodness of fit hypothesis at a certain significance level α if

$$C_{\mathbf{k}} > \chi^2_{1-\alpha}(n - \text{rank}(A)),$$

where $\chi^2_{1-\alpha}(m)$ is the $(1-\alpha)$-quantile of the distribution $\chi^2(m)$.

Let $\mathcal{Y}$ be the set k observations with the same MLE as $\mathcal{U}_0$; it is known that every sample in $\mathcal{Y}$ has the same probability under our independence model [39]; therefore exact methods based on the MLE-statistics consist in estimating the exact p-value, that is the ratio of cardinalities:

$$p_{\mathcal{U}_0} = \frac{|\{\mathcal{U} \in \mathcal{Y} \mid C_{\mathcal{U}} \geq C_{\mathcal{U}_0}\}|}{|\mathcal{Y}|} \tag{5.10}$$

Since the order of the sets involved is in general very large, Monte Carlo techniques are used in order to randomly walking through the set $\mathcal{Y}$. As observed in [9,40], this aim can be accomplished by reducing to the problem of sampling in the fiber $\mathcal{F}(\mathbf{k}_0)$, as we explain below.

Let $\mathcal{F} = \mathcal{F}(\mathbf{k}_0)$ be the fiber of $\mathbf{k}_0$ in the lattice $\mathcal{L} = \mathcal{L}_A$. Observe that for every $\mathbf{h} = (h_1, ..., h_n) \in \mathcal{F}$ there are exactly $\frac{k!}{h_1! \cdot \ldots \cdot h_n!}$ elements of $\mathcal{Y}$ having $\mathbf{h}$ as their frequency vector. Therefore the uniform distribution on $\mathcal{Y}$ induces on $\mathcal{F}$ the hypergeometric distribution, such that the probability of $\mathbf{h}$ is proportional to $\tau(\mathbf{h}) = \frac{1}{h_1! \cdot \ldots \cdot h_n!}$. Then the problem of uniformly sampling from $\mathcal{Y}$ is reduced to that of sampling in the fiber $\mathcal{F}$ according to the hypergeometric distribution, and this can be performed by using the Diaconis-Sturmfels algorithm. Therefore an approximation of the p-value (5.10) can be obtained by running the following algorithm:

Diaconis-Sturmfels exact goodness-of-fit test

Let $\mathcal{B}$ be a Markov basis for $\mathcal{L}$. Fix a (large) number N of iterations.

Start from $\mathbf{h}_0 = \mathbf{k}_0 \in \mathcal{F}$; introduce a counter function T and initialize $T(0) = 0$.

Step 1. Randomly choose $\mathbf{b} \in \mathcal{B}$ and $\epsilon \in \{\pm 1\}$.

Step 2. If $\mathbf{h}_r + \epsilon\mathbf{b} \notin \mathbf{N}^n$ then set $\mathbf{h}_{r+1} = \mathbf{h}_r$; else, move to $\mathbf{h}_{r+1} = \mathbf{h}_r + \epsilon\mathbf{b}$ with probability $\min\left\{\frac{\tau(\mathbf{h}_r+\epsilon\mathbf{b})}{\tau(\mathbf{h}_r)}, 1\right\}$, i.e. put

$$\mathbf{h}_{r+1} = \begin{cases} \mathbf{h}_r + \epsilon\mathbf{b} & \text{with probability } \mu = \min\left\{\frac{\tau(\mathbf{h}_r+\epsilon\mathbf{b})}{\tau(\mathbf{h}_r)}, 1\right\} \\ \mathbf{h}_r & \text{with probability } 1-\mu \end{cases}$$

Step 3. Compute the Pearson coefficient $C_{\mathbf{h}_{r+1}}$.

Step 4. Put

$$T(r+1) = \begin{cases} T(r)+1 & \text{if } C_{\mathbf{h}_{r+1}} \geq C_{\mathbf{k}_0} \\ T(r) & \text{if } C_{\mathbf{h}_{r+1}} < C_{\mathbf{k}_0} \end{cases}$$

Step 5. If $r = N$ then stop and output $IC = \frac{T(N)}{N}$; else put $r := r+1$ and go to Step 1.

The output $IC(\mathbf{k}_0)$ will be called the **Monte Carlo p-value** of $\mathbf{k}_0$ (or $\mathcal{U}_0$). The goodness-of-fit hypothesis is rejected if $IC(\mathbf{k}_0) < \alpha$ for a significance level α.

TABLE 1. Polymorphisms included in the analyses

Gene	**Function**	**Polymorphism**	**rs**	**Genotype frequency (%)**		
DNA repair				w/w	w/m	m/m
1 *ERCC2/XPD*	NER	Asp312Asn	rs1799793	38.4	45.3	16.2
2 *ERCC2/XPD*	NER	Lys751Gln	rs13181	35.7	46.1	18.2
3 *PCNA*	BER	6084 G > C(30-UTR)	rs3626	80	18.9	1.1
4 *XRCC1*	BER	Arg194Trp	rs1799782	87.6	12.2	0.3
5 *XRCC1*	BER	Pro206Pro	rs915927	30.9	46	23.1
6 *XRCC1*	BER	Arg399Gln	rs25487	44.8	43.6	11.6
7 *XRCC3*	DSBR	17893 A > G (IVS6-14)	rs1799796	51.1	40.1	8.9
8 *XRCC3*	DSBR	Thr241Met	rs861539	34.6	49.8	15.6
9 *APE1*	BER	Asp148Glu	rs3136820	29.3	48.2	22.5
10 *ERCC1*	NER	Asn118Asn	rs3177700	35.6	46.8	17.6
11 *MGMT*	DRR	Leu84Phe	rs12917	73.6	24.6	1.8
12 *hOGG1*	BER	Ser326Cys	rs1052133	62.6	32	5.4
13 *BRCA1*	DSBR	Pro871Leu	rs799917	42.5	47.7	9.8
14 *BRCA2*	DSBR	Asn372His	rs144848	51.6	41.3	7.1
15 *NBS1*	DSBR	Glu185Gln	rs1805794	49.2	40.2	10.6
16 *RAD51*	DSBR	135 G > C (5' -UTR)	rs1801320	87	12.7	0.3
17 *RAD51*	DSBR	172 G > T (5'-UTR)	rs1801321	33.8	48.7	17.5
18 *RAD52*	DSBR	2259 C > T (3'-UTR)	rs11226	33.7	46.8	19.4
19 *XRCC2*	DSBR	Arg188His	rs3218536	83.2	16	0.9
20 *LIG4*	DSBR	Ala3Val	rs1805389	88.7	7.7	3.5
21 *LIG4*	DSBR	Thr9Ile	rs1805388	72.1	25	2.8
22 *TP53*	Cell cycle/apoptosis	Arg72Pro	rs1042522	57.5	36.5	5.9
Metabolic						
23 *MnSOD*	Oxidative scavenger	Ala9Val	rs4880	25.2	52.7	22.1
24 *NQO1*	Oxidative scavenger	Pro187Ser	rs1800566	65	31.4	3.6
25 *COMT*	Phase 1	Val158Met	rs4680	24.8	52.1	23.1
26 *MPO*	Phase 1	G > A SP1 site	rs2333227	60.3	35.4	4.2
27 *SULT1A1*	Phase 2	Arg213His	rs9282861	44.7	43.5	11.8
28 *GSTM3*	Phase 2	3 bp deletion (*A,*B)		68.7	28.4	3
29 *GSTP1*	Phase 2	Ile105Val	rs1695	44.1	44.1	11.8
30 *GSTP1*	Phase 2	Ala114Val	rs1138272	83.9	15	1.1
31 *CYP1A1*	Phase 1	Ile462Val	rs1048943	83.6	15.4	1
32 *CYP1B1*	Phase 1	Val432Leu	rs1056836	16.4	48.6	35
				w/w	w/m + m/m	
33 *GSTT1*	Phase 2	Gene deletion (*1, *2/*2)		75.5	24.5	
34 *GSTM1*	Phase 2	Gene deletion (*2/*2, *1)		55.6	44.4	
35 *NAT2*	Phase 2	Slow/rapid acetylator		42.7	57.3	

Table 2. Interaction results for bladder cancer

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	0,71	0,56	0,89	0,95	0,92	0,56	0,43	0,81	0,52	0,8	0,57	0,82	0,75	0,66	0,62	0,76	0,99	0,12
2	0,56	0,34	0,5	0,68	0,41	0,62	0,53	0,31	0,58	0,9	0,59	0,94	0,8	0,2	0,61	0,44	0,94	0,89
3	0,89	0,5	0,32	0,62	0,87	0,81	0,69	0,92	0,71	0,52	0,61	0,38	0,88	0,42	0,66	0,76	0,86	0,55
4	0,95	0,68	0,62	0,92	0,89	0,89	0,32	0,63	0,71	0,89	0,63	0,79	0,95	0,23	0,2	0,84	0,55	0,12
5	0,92	0,41	0,87	0,89	0,54	0,33	0,13	0,82	0,72	0,71	0,62	0,96	0,65	0,42	0,71	0,62	0,81	0,68
6	0,56	0,62	0,81	0,89	0,33	0,75	**0,032**	1	0,63	0,65	0,26	0,99	0,18	0,28	0,41	0,76	0,41	0,71
7	0,43	0,53	0,69	0,32	0,13	**0,032**	0,53	0,74	0,67	0,82	0,49	0,71	0,14	0,11	0,48	0,35	0,54	0,38
8	0,81	0,31	0,92	0,63	0,82	1	0,74	0,94	0,91	0,66	0,49	0,84	0,61	0,32	0,93	0,45	0,59	0,31
9	0,52	0,58	0,71	0,71	0,72	0,63	0,67	0,91	0,42	0,86	0,26	0,8	0,79	0,17	0,27	0,47	0,9	0,44
10	0,8	0,9	0,52	0,89	0,71	0,65	0,82	0,66	0,86	0,44	0,53	0,6	0,47	0,29	0,78	0,22	0,83	0,8
11	0,57	0,59	0,61	0,63	0,62	0,26	0,49	0,49	0,26	0,53	0,18	0,76	0,55	0,11	0,65	0,48	0,56	0,29
12	0,82	0,94	0,38	0,79	0,96	0,99	0,71	0,84	0,8	0,6	0,76	0,89	0,95	0,41	0,87	0,72	0,89	0,5
13	0,75	0,8	0,88	0,95	0,65	0,18	0,14	0,61	0,79	0,47	0,55	0,95	0,66	0,35	0,88	0,23	0,83	0,7
14	0,66	0,2	0,42	0,23	0,42	0,28	0,11	0,32	0,17	0,29	0,11	0,41	0,35	0,12	0,35	0,48	0,75	0,52
15	0,62	0,61	0,66	0,2	0,71	0,41	0,48	0,93	0,27	0,78	0,65	0,87	0,88	0,35	0,53	0,64	0,4	0,11
16	0,76	0,44	0,76	0,84	0,62	0,76	0,35	0,45	0,47	0,22	0,48	0,72	0,23	0,48	0,64	0,9	0,96	0,35
17	0,99	0,94	0,86	0,55	0,81	0,41	0,54	0,59	0,9	0,83	0,56	0,89	0,83	0,75	0,4	0,96	0,7	0,85
18	0,12	0,89	0,55	0,12	0,68	0,71	0,38	0,31	0,44	0,8	0,29	0,5	0,7	0,52	0,11	0,35	0,85	0,4
19	0,96	0,81	0,87	0,68	0,85	0,9	0,55	0,54	0,56	0,79	0,77	0,74	0,85	0,24	0,88	0,99	0,85	0,66
20	0,63	0,17	0,85	0,94	0,79	0,95	0,78	0,42	0,85	0,32	0,69	0,59	0,77	0,41	0,88	0,92	0,73	0,26
21	0,73	0,82	0,15	0,36	0,63	0,7	0,13	0,11	0,47	0,61	0,78	0,66	0,87	0,16	0,53	0,95	0,074	0,43
22	0,44	0,52	0,68	0,76	0,73	0,89	0,57	0,41	0,4	0,69	0,36	0,66	0,58	0,16	0,67	0,7	0,88	0,34
23	0,39	0,45	0,65	0,48	0,97	0,64	0,63	0,63	0,93	0,6	0,75	0,56	0,65	0,59	0,72	0,9	0,45	0,47
24	0,27	0,42	0,48	0,28	0,02	0,19	0,51	0,21	0,32	0,38	0,03	0,84	0,68	0,22	0,83	0,56	0,93	0,89
25	0,21	0,45	0,33	0,34	0,1	0,16	0,14	0,13	**0,009**	0,2	**0,039**	0,46	0,91	0,48	0,39	0,89	0,91	0,48
26	0,47	0,69	0,76	0,53	0,88	0,88	0,41	0,93	0,11	0,79	0,22	0,9	0,41	0,15	0,34	0,82	0,7	0,43
27	0,63	0,14	0,87	0,95	0,91	0,86	0,79	0,79	0,88	0,97	0,64	0,97	0,9	0,43	0,96	0,91	0,96	0,93
28	0,94	0,8	0,43	0,35	0,51	0,59	0,4	0,86	0,058	0,76	0,3	0,91	0,69	0,3	0,6	0,93	0,66	0,55
29	0,22	0,76	0,61	0,88	0,69	0,38	0,6	0,83	0,07	0,58	0,74	0,06	0,98	**0,014**	0,79	0,94	0,98	0,8
30	0,9	0,51	0,76	0,73	0,55	0,77	0,74	0,93	0,71	0,58	0,44	0,78	0,61	0,12	0,77	0,53	0,84	0,69
31	0,9	0,48	0,5	0,23	0,27	0,11	0,85	0,86	0,21	0,67	0,21	0,88	0,46	0,19	0,49	0,67	0,25	0,44
32	0,8	0,71	0,24	0,78	**0,025**	0,51	0,92	0,73	0,64	0,54	0,56	0,75	0,29	0,14	0,31	0,22	0,7	0,3
33	0,29	0,21	0,63	0,84	0,44	0,6	0,77	0,86	0,4	0,092	0,075	0,79	0,51	0,23	0,088	0,17	0,13	0,36
34	0,31	**0,048**	0,23	0,34	0,31	0,19	0,22	0,23	0,2	0,16	0,18	0,42	0,32	0,1	0,13	0,17	0,45	0,15
35	0,7	0,4	0,8	0,93	0,51	0,081	0,83	0,99	0,8	0,86	0,55	0,72	0,86	0,48	0,78	0,69	0,98	0,83

	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
1	0,96	0,63	0,73	0,44	0,39	0,27	0,21	0,47	0,63	0,94	0,22	0,9	0,9	0,8	0,29	0,31	0,7
2	0,81	0,17	0,82	0,52	0,45	0,42	0,45	0,69	0,14	0,8	0,76	0,51	0,48	0,71	0,21	**0,048**	0,4
3	0,87	0,85	0,15	0,68	0,65	0,48	0,33	0,76	0,87	0,43	0,61	0,76	0,5	0,24	0,63	0,23	0,8
4	0,68	0,94	0,36	0,76	0,48	0,28	0,34	0,53	0,95	0,35	0,88	0,73	0,23	0,78	0,84	0,34	0,93
5	0,85	0,79	0,63	0,73	0,97	0,02	0,1	0,88	0,91	0,51	0,69	0,55	0,27	**0,025**	0,44	0,31	0,51
6	0,9	0,95	0,7	0,89	0,64	0,19	0,16	0,88	0,86	0,59	0,38	0,77	0,11	0,51	0,6	0,19	0,081
7	0,55	0,78	0,13	0,57	0,63	0,51	0,14	0,41	0,79	0,4	0,6	0,74	0,85	0,92	0,77	0,22	0,83
8	0,54	0,42	0,11	0,41	0,63	0,21	0,13	0,93	0,79	0,86	0,83	0,93	0,86	0,73	0,86	0,23	0,99
9	0,56	0,85	0,47	0,4	0,93	0,32	**0,009**	0,11	0,88	0,058	0,07	0,71	0,21	0,64	0,4	0,2	0,8
10	0,79	0,32	0,61	0,69	0,6	0,38	0,2	0,79	0,97	0,76	0,58	0,58	0,67	0,54	0,092	0,16	0,86
11	0,77	0,69	0,78	0,36	0,75	0,03	**0,039**	0,22	0,64	0,3	0,74	0,44	0,21	0,56	0,075	0,18	0,55
12	0,74	0,59	0,66	0,66	0,56	0,84	0,46	0,9	0,97	0,91	0,06	0,78	0,88	0,75	0,79	0,42	0,72
13	0,85	0,77	0,87	0,58	0,65	0,68	0,91	0,41	0,9	0,69	0,98	0,61	0,46	0,29	0,51	0,32	0,86
14	0,24	0,41	0,16	0,16	0,59	0,22	0,48	0,15	0,43	0,3	0,014	0,12	0,19	0,14	0,23	0,1	0,48
15	0,88	0,88	0,53	0,67	0,72	0,83	0,39	0,34	0,96	0,6	0,79	0,77	0,49	0,31	0,088	0,13	0,78
16	0,99	0,92	0,95	0,7	0,9	0,56	0,89	0,82	0,91	0,93	0,94	0,53	0,67	0,22	0,17	0,17	0,69
17	0,85	0,73	0,074	0,88	0,45	0,93	0,91	0,7	0,96	0,66	0,98	0,84	0,25	0,7	0,13	0,45	0,98
18	0,66	0,26	0,43	0,34	0,47	0,89	0,48	0,43	0,93	0,55	0,8	0,69	0,44	0,3	0,36	0,15	0,83
19	0,92	0,98	0,61	0,63	0,65	0,69	0,82	0,91	0,51	0,47	1	0,1	0,59	0,12	0,39	0,15	0,97
20	0,98	0,68	0,67	0,71	0,85	0,67	0,77	0,44	0,98	0,95	0,86	0,58	0,3	0,41	0,37	**0,049**	0,41
21	0,61	0,67	0,49	0,53	0,64	0,21	0,68	0,33	0,98	0,73	0,74	0,46	0,37	0,66	0,26	**0,0099**	0,31
22	0,63	0,71	0,53	0,28	0,44	0,5	0,49	0,66	0,097	0,28	0,8	0,26	0,23	0,67	0,5	0,2	0,36
23	0,65	0,85	0,64	0,44	0,72	0,63	0,18	0,84	0,7	0,96	0,57	0,32	0,1	0,73	0,88	0,35	0,56
24	0,69	0,67	0,21	0,5	0,63	0,16	0,11	0,52	0,69	0,79	0,81	0,15	0,35	0,42	0,17	0,17	0,63
25	0,82	0,77	0,68	0,49	0,18	0,11	0,061	0,1	0,54	0,1	0,27	0,06	0,12	0,35	0,2	0,062	0,3
26	0,91	0,44	0,33	0,66	0,84	0,52	0,1	0,38	0,84	0,59	0,32	0,33	0,57	0,21	0,53	0,072	0,56
27	0,51	0,98	0,98	0,097	0,7	0,69	0,54	0,84	0,77	0,42	0,29	1	0,76	0,6	0,42	0,35	0,66
28	0,47	0,95	0,73	0,28	0,96	0,79	0,1	0,59	0,42	0,58	0,8	0,87	0,17	0,76	0,13	0,13	0,82
29	1	0,86	0,74	0,8	0,57	0,81	0,27	0,32	0,29	0,8	0,58	0,68	0,24	0,35	0,66	0,33	0,83
30	0,1	0,58	0,46	0,26	0,32	0,15	0,06	0,33	1	0,87	0,68	0,41	0,48	0,89	0,62	0,1	0,9
31	0,59	0,3	0,37	0,23	0,1	0,35	0,12	0,57	0,76	0,17	0,24	0,48	0,29	0,18	0,6	**0,038**	0,71
32	0,12	0,41	0,66	0,67	0,73	0,42	0,35	0,21	0,6	0,76	0,35	0,89	0,18	0,34	0,57	0,18	0,35
33	0,39	0,37	0,26	0,5	0,88	0,17	0,2	0,53	0,42	0,13	0,66	0,62	0,6	0,57	0,27	0,081	0,4
34	0,15	**0,049**	**0,0099**	0,2	0,35	0,17	0,062	0,072	0,35	0,13	0,33	0,1	**0,038**	0,18	0,081	**0,029**	0,13
35	0,97	0,41	0,31	0,36	0,56	0,63	0,3	0,56	0,66	0,82	0,83	0,9	0,71	0,35	0,4	0,13	0,65

Table 3. Interaction results for lung cancer

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	0,69	0,48	0,31	0,11	0,93	0,23	0,4	0,64	1	0,7	0,9	0,35	0,73	0,31	0,44	0,51	0,57	0,72
2	0,48	0,47	0,87	0,26	0,51	0,29	0,71	0,55	0,85	0,34	0,65	0,45	0,9	0,46	0,77	0,65	0,59	0,74
3	0,31	0,87	0,5	0,22	0,25	0,096	0,99	0,94	0,97	0,38	0,93	0,52	0,98	0,12	0,36	0,54	0,97	0,85
4	0,11	0,26	0,22	0,13	0,22	**0,012**	0,34	0,37	0,2	0,079	0,39	0,23	0,13	0,059	0,24	0,12	0,28	0,3
5	0,93	0,51	0,25	0,22	0,4	0,14	0,18	0,2	0,85	**0,041**	0,69	0,42	0,68	0,014	0,29	0,12	0,7	0,21
6	0,23	0,29	0,096	**0,012**	0,14	**0,027**	0,066	**0,023**	0,16	**0,039**	0,21	0,069	**0,048**	**0,00059**	**0,036**	0,02	0,22	**0,018**
7	0,4	0,71	0,99	0,34	0,18	0,066	0,66	0,84	0,87	0,49	0,8	0,58	0,35	0,17	0,68	0,059	0,74	0,81
8	0,64	0,55	0,94	0,37	0,2	**0,023**	0,84	0,68	0,36	0,72	0,74	0,79	0,74	0,22	0,65	0,24	0,8	0,61
9	1	0,85	0,97	0,2	0,85	0,16	0,87	0,36	0,91	0,43	0,28	0,78	0,86	0,13	0,37	0,63	0,78	0,92
10	0,7	0,34	0,38	0,079	**0,041**	**0,039**	0,49	0,72	0,43	0,15	0,48	0,27	0,61	0,054	0,18	0,28	0,12	0,16
11	0,9	0,65	0,93	0,39	0,69	0,21	0,8	0,74	0,28	0,48	0,46	0,59	0,97	0,23	0,81	0,43	0,45	0,61
12	0,35	0,45	0,52	0,23	0,42	0,069	0,58	0,79	0,78	0,27	0,59	0,24	0,62	0,11	0,8	0,43	0,87	0,49
13	0,73	0,9	0,98	0,13	0,68	**0,048**	0,35	0,74	0,86	0,61	0,97	0,62	0,81	0,12	0,98	0,23	0,65	0,24
14	0,31	0,46	0,12	0,059	**0,014**	**0,00059**	0,17	0,22	0,13	0,054	0,23	0,11	0,12	0,025	0,19	0,056	0,24	0,22
15	0,44	0,77	0,36	0,24	0,29	**0,036**	0,68	0,65	0,37	0,18	0,81	0,8	0,98	0,19	0,51	0,4	0,43	0,22
16	0,51	0,65	0,54	0,12	0,12	**0,02**	0,059	0,24	0,63	0,28	0,43	0,43	0,23	0,056	0,4	0,17	0,28	0,48
17	0,57	0,59	0,97	0,28	0,7	0,22	0,74	0,8	0,78	0,12	0,45	0,87	0,65	0,24	0,43	0,28	0,56	0,51
18	0,72	0,74	0,85	0,3	0,21	0,018	0,81	0,61	0,92	0,16	0,61	0,49	0,24	0,22	0,22	0,48	0,51	0,33
19	0,78	0,52	0,42	0,45	0,87	0,17	0,24	0,17	0,43	0,087	0,93	0,91	0,76	0,11	0,93	0,77	0,87	0,7
20	0,58	0,97	0,96	0,45	0,67	0,17	0,26	0,59	0,79	0,78	0,81	0,33	0,81	0,18	0,5	0,46	0,83	0,9
21	0,27	0,85	0,92	0,27	0,64	**0,047**	0,34	0,79	0,89	0,65	0,5	0,88	0,52	0,12	0,74	0,74	0,8	0,47
22	0,29	0,37	0,073	0,14	**0,019**	**0,023**	0,37	0,062	0,33	0,076	0,19	0,36	**0,011**	0,041	0,23	0,18	0,67	0,54
23	0,65	0,37	0,68	0,39	0,28	0,27	0,52	0,36	0,71	0,54	0,64	0,46	0,85	0,32	0,097	0,41	0,26	0,68
24	**0,033**	**0,026**	**0,0052**	**0,00077**	**0,0035**	**0,002**	**0,021**	**0,0085**	**0,046**	**0,00026**	**0,011**	**0,028**	**0,05**	**0,0021**	0,13	**0,0087**	**0,022**	0,17
25	0,43	0,45	0,7	**0,034**	0,65	0,2	0,98	0,53	0,67	0,62	0,71	0,55	0,82	0,11	0,99	0,73	0,97	0,59
26	1	0,8	0,94	0,24	0,43	0,21	0,36	0,85	1	0,7	0,95	0,77	0,12	0,11	0,24	0,72	0,1	0,84
27	0,98	0,93	0,95	0,45	0,58	0,16	0,96	0,95	0,86	0,41	0,93	0,81	0,58	0,51	0,97	0,64	0,81	0,56
28	0,77	0,5	0,32	0,24	0,44	0,14	0,81	0,73	0,64	0,34	0,78	0,45	0,21	0,02	0,5	0,081	**0,027**	0,11
29	0,62	0,58	0,51	0,35	0,42	0,1	0,31	0,37	0,63	0,16	0,78	0,3	0,63	0,28	0,79	0,67	0,76	0,69
30	0,71	0,35	0,94	0,61	0,68	0,24	0,97	0,71	0,85	0,5	0,61	0,58	0,42	0,067	0,58	0,12	0,66	0,31
31	0,41	0,4	0,71	0,3	0,58	0,086	0,83	0,54	0,8	0,21	0,78	0,78	0,77	0,07	0,85	0,85	0,44	0,27
32	0,94	0,98	0,69	0,29	0,86	0,23	0,51	0,99	1	0,42	0,55	0,22	0,78	0,11	0,99	0,35	0,81	0,31
33	0,94	0,89	0,91	0,4	0,8	0,12	0,29	0,88	0,85	0,4	0,88	0,58	0,82	0,072	0,57	0,61	0,23	0,57
34	0,94	0,83	0,95	0,2	0,29	0,089	0,51	0,96	1	0,22	0,65	0,51	0,97	0,12	0,88	0,46	0,42	0,72
35	0,79	0,38	0,88	0,26	0,81	0,17	0,64	0,42	0,96	0,18	0,37	0,66	0,58	0,11	0,8	0,25	0,15	0,58

19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
0,78	0,58	0,27	0,29	0,65	**0,033**	0,43	1	0,98	0,77	0,62	0,71	0,41	0,94	0,94	0,94	0,79
0,52	0,97	0,85	0,37	0,37	**0,026**	0,45	0,8	0,93	0,5	0,58	0,35	0,4	0,98	0,89	0,83	0,38
0,42	0,96	0,92	0,073	0,68	**0,0052**	0,7	0,94	0,95	0,32	0,51	0,94	0,71	0,69	0,91	0,95	0,88
0,45	0,45	0,27	0,14	0,39	**0,00077**	**0,034**	0,24	0,45	0,24	0,35	0,61	0,3	0,29	0,4	0,2	0,26
0,87	0,67	0,64	**0,019**	0,28	**0,0035**	0,65	0,43	0,58	0,44	0,42	0,68	0,58	0,86	0,8	0,29	0,81
0,17	0,17	**0,047**	**0,023**	0,27	**0,002**	0,2	0,21	0,16	0,14	0,1	0,24	0,086	0,23	0,12	0,089	0,17
0,24	0,26	0,34	0,37	0,52	**0,021**	0,98	0,36	0,96	0,81	0,31	0,97	0,83	0,51	0,29	0,51	0,64
0,17	0,59	0,79	0,062	0,36	**0,0085**	0,53	0,85	0,95	0,73	0,37	0,71	0,54	0,99	0,88	0,96	0,42
0,43	0,79	0,89	0,33	0,71	**0,046**	0,67	1	0,86	0,64	0,63	0,85	0,8	1	0,85	1	0,96
0,087	0,78	0,65	0,076	0,54	**0,00026**	0,62	0,7	0,41	0,34	0,16	0,5	0,21	0,42	0,4	0,22	0,18
0,93	0,81	0,5	0,19	0,64	**0,011**	0,71	0,95	0,93	0,78	0,78	0,61	0,78	0,55	0,88	0,65	0,37
0,91	0,33	0,88	0,36	0,46	**0,028**	0,55	0,77	0,81	0,45	0,3	0,58	0,78	0,22	0,58	0,51	0,66
0,76	0,81	0,52	**0,011**	0,85	**0,05**	0,82	0,12	0,58	0,21	0,63	0,42	0,77	0,78	0,82	0,97	0,58
0,11	0,18	0,12	**0,041**	0,32	**0,0021**	0,11	0,11	0,51	0,02	0,28	0,067	0,07	0,11	0,072	0,12	0,11
0,93	0,5	0,74	0,23	0,097	0,13	0,99	0,24	0,97	0,5	0,79	0,58	0,85	0,99	0,57	0,88	0,8
0,77	0,46	0,74	0,18	0,41	**0,0087**	0,73	0,72	0,64	0,081	0,67	0,12	0,85	0,35	0,61	0,46	0,25
0,87	0,83	0,8	0,67	0,26	**0,022**	0,97	0,1	0,81	**0,027**	0,76	0,66	0,44	0,81	0,23	0,42	0,15
0,7	0,9	0,47	0,54	0,68	0,17	0,59	0,84	0,56	0,11	0,69	0,31	0,27	0,31	0,57	0,72	0,58
0,74	0,87	0,57	0,3	0,54	**0,0086**	0,99	0,97	0,96	0,52	0,81	0,42	0,96	0,84	0,91	0,97	0,72
0,87	0,46	0,58	0,22	0,66	**0,039**	0,79	0,68	0,82	0,2	0,96	0,67	0,93	0,88	0,85	0,65	0,63
0,57	0,58	0,75	0,34	0,66	**0,037**	0,99	0,78	0,28	0,11	0,91	0,56	0,91	0,24	0,93	0,73	0,87
0,3	0,22	0,34	0,07	0,054	**0,02**	0,62	0,36	**0,023**	0,067	**0,00035**	0,13	0,086	0,69	0,19	0,15	0,17
0,54	0,66	0,66	0,054	0,6	**0,04**	0,89	0,6	0,92	0,38	0,36	0,28	0,52	0,89	0,97	0,6	0,53
0,0086	**0,039**	**0,037**	**0,02**	**0,04**	**0,002**	**0,0084**	**0,043**	**0,0092**	**0,014**	**0,028**	**0,0096**	**0,0077**	**0,016**	**0,013**	**0,021**	**0,016**
0,99	0,79	0,99	0,62	0,89	**0,0084**	0,89	0,85	0,9	0,81	0,37	1	0,85	0,68	0,54	0,87	0,98
0,97	0,68	0,78	0,36	0,6	**0,043**	0,85	0,98	0,99	0,59	0,33	0,98	0,72	0,66	0,98	0,86	0,57
0,96	0,82	0,28	**0,023**	0,92	**0,0092**	0,9	0,99	0,92	0,9	0,35	0,84	0,51	0,83	0,6	1	0,96
0,52	0,2	0,11	0,067	0,38	**0,014**	0,81	0,59	0,9	0,27	0,6	0,81	0,48	0,71	0,44	0,41	0,39
0,81	0,96	0,91	**0,00035**	0,36	**0,028**	0,37	0,33	0,35	0,6	0,28	0,63	0,081	**0,036**	0,72	0,78	0,68
0,42	0,67	0,56	0,13	0,28	**0,0096**	1	0,98	0,84	0,81	0,63	0,78	0,49	0,92	0,93	0,97	0,61
0,96	0,93	0,91	0,086	0,52	**0,0077**	0,85	0,72	0,51	0,48	0,081	0,49	0,41	0,62	0,71	0,78	0,78
0,84	0,88	0,24	0,69	0,89	**0,016**	0,68	0,66	0,83	0,71	**0,036**	0,92	0,62	0,81	0,97	0,9	0,89
0,91	0,85	0,93	0,19	0,97	**0,013**	0,54	0,98	0,6	0,44	0,72	0,93	0,71	0,97	0,62	0,82	0,4
0,97	0,65	0,73	0,15	0,6	**0,021**	0,87	0,86	1	0,41	0,78	0,97	0,78	0,9	0,82	0,91	0,12
0,72	0,63	0,87	0,17	0,53	**0,016**	0,98	0,57	0,96	0,39	0,68	0,61	0,78	0,89	0,4	0,12	0,54

Table 4. Interaction results for leukaemia

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	0,56	0,49	0,75	0,78	0,78	0,51	0,61	0,32	0,36	0,22	0,73	0,49	0,61	0,87	0,91	0,21	0,082	0,39
2	0,49	0,21	0,29	0,6	0,35	0,22	0,3	0,19	0,32	0,7	0,75	0,27	0,19	0,43	0,61	0,17	**0,0072**	0,34
3	0,75	0,29	0,74	0,85	0,26	0,22	0,97	0,26	0,64	0,94	0,6	0,95	0,66	0,99	0,84	0,47	0,41	0,94
4	0,78	0,6	0,85	0,41	0,58	**0,048**	0,83	0,13	0,35	0,91	0,58	0,34	0,17	0,4	0,66	0,23	0,21	0,18
5	0,78	0,35	0,26	0,58	0,13	0,33	0,71	0,11	0,22	0,4	0,77	0,66	0,45	0,33	0,34	**0,021**	**0,026**	0,39
6	0,51	0,22	0,22	**0,048**	0,33	0,076	0,46	**0,029**	**0,016**	0,26	**0,016**	0,63	0,46	0,2	**0,017**	0,16	0,3	0,66
7	0,61	0,3	0,97	0,83	0,71	0,46	0,86	0,15	0,55	0,6	0,82	0,49	0,57	0,23	0,63	0,61	0,37	0,54
8	0,32	0,19	0,26	0,13	0,11	**0,029**	0,15	0,065	0,15	0,31	0,27	0,34	0,32	0,71	0,83	0,45	0,16	0,87
9	0,36	0,32	0,64	0,35	0,22	**0,016**	0,55	0,15	0,28	0,92	0,25	**0,011**	0,21	0,8	0,56	0,61	**0,011**	0,38
10	0,22	0,7	0,94	0,91	0,4	0,26	0,6	0,31	0,92	0,76	0,32	0,6	0,42	0,34	0,99	0,14	0,59	0,82
11	0,73	0,75	0,6	0,58	0,77	**0,016**	0,82	0,27	0,25	0,32	0,61	0,96	0,7	0,85	0,95	0,81	0,56	0,96
12	0,49	0,27	0,95	0,34	0,66	0,63	0,49	0,34	**0,011**	0,6	0,96	0,49	0,25	0,89	0,18	0,11	0,37	0,35
13	0,61	0,19	0,66	0,17	0,45	0,46	0,57	0,32	0,21	0,42	0,7	0,25	0,16	0,64	0,057	0,19	0,23	0,18
14	0,87	0,43	0,99	0,4	0,33	0,2	0,23	0,71	0,8	0,34	0,85	0,89	0,64	0,89	0,68	0,59	0,34	0,99
15	0,91	0,61	0,84	0,66	0,34	**0,017**	0,63	0,83	0,56	0,99	0,95	0,18	0,057	0,68	0,87	0,3	0,11	0,99
16	0,21	0,17	0,47	0,23	**0,021**	0,16	0,61	0,45	0,61	0,14	0,81	0,11	0,19	0,59	0,3	0,19	0,2	0,63
17	0,082	**0,0072**	0,41	0,21	**0,026**	0,3	0,37	0,16	**0,011**	0,59	0,56	0,37	0,23	0,34	0,11	0,2	0,095	0,6
18	0,39	0,34	0,94	0,18	0,39	0,66	0,54	0,87	0,38	0,82	0,96	0,35	0,18	0,99	0,99	0,63	0,6	0,86
19	0,48	0,27	0,94	0,45	0,32	0,65	0,9	0,63	0,66	0,74	1	0,085	0,066	0,79	1	0,75	0,69	0,56
20	0,47	0,17	0,89	0,31	0,61	0,14	0,86	0,58	0,88	0,91	0,82	0,13	0,21	0,84	0,78	0,053	0,24	0,39
21	0,51	**0,0095**	0,79	0,27	0,68	0,19	0,62	0,98	0,72	0,59	0,73	0,4	0,21	0,94	1	0,95	0,057	0,96
22	1	0,27	0,81	0,56	0,39	0,44	0,82	0,53	0,35	0,96	1	0,64	0,37	0,16	1	0,81	0,29	0,82
23	0,094	0,11	0,26	0,069	0,15	**0,044**	0,49	**0,022**	0,25	0,21	0,44	0,22	0,11	0,51	0,11	0,02	0,08	0,28
24	0,12	0,83	0,14	0,37	0,47	0,28	0,78	0,49	0,71	0,24	0,92	0,88	0,19	0,84	0,55	0,11	0,12	0,8
25	0,87	0,53	0,94	0,75	0,63	0,28	0,8	0,23	0,55	0,64	0,26	0,24	0,68	0,9	0,6	0,58	**0,023**	0,91
26	0,94	0,42	0,8	0,8	0,82	0,45	0,94	0,42	0,2	0,58	0,87	0,88	0,071	0,9	1	0,69	0,4	0,94
27	0,74	0,51	0,98	0,79	0,41	0,12	0,94	0,63	0,58	0,83	0,58	0,11	0,78	0,98	0,56	0,58	0,21	0,93
28	0,98	0,38	0,99	0,66	0,086	0,27	0,72	0,69	0,52	0,9	0,89	0,65	0,46	1	0,38	0,32	0,53	1
29	0,87	0,76	0,92	0,37	0,28	0,57	0,92	0,38	0,54	0,96	0,64	0,36	0,44	0,22	0,63	0,55	0,27	0,31
30	0,65	0,25	0,62	0,57	0,33	0,18	0,85	0,18	**0,0036**	0,55	0,36	0,3	0,39	0,47	0,91	0,55	0,22	0,34
31	0,11	0,052	0,058	0,01	**0,015**	**0,0063**	0,06	**0,0019**	0,01	0,078	0,11	**0,03**	**0,008**	0,083	**0,0032**	**0,01**	**0,0065**	**0,0018**
32	0,3	0,07	0,073	0,069	0,061	**0,029**	0,4	0,1	0,19	0,48	0,27	0,27	0,68	0,91	0,82	0,61	**0,082**	0,51
33	0,96	0,61	0,97	0,66	0,65	0,36	0,94	0,084	0,51	0,25	0,94	0,52	0,4	0,31	0,74	0,47	0,43	0,54
34	0,35	0,21	0,97	0,54	0,64	0,4	0,68	0,15	0,58	0,52	0,5	0,79	0,39	0,99	0,21	0,65	0,42	0,91
35	0,56	0,46	0,5	0,35	0,42	0,1	0,6	0,1	0,38	0,33	0,19	0,32	0,26	0,75	0,13	0,15	0,26	0,63

	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
1	0,48	0,47	0,51	1	0,094	0,12	0,87	0,94	0,74	0,98	0,87	0,65	0,11	0,3	0,96	0,35	0,56
2	0,27	0,17	**0,0095**	0,27	0,11	0,83	0,53	0,42	0,51	0,38	0,76	0,25	0,052	0,07	0,61	0,21	0,46
3	0,94	0,89	0,79	0,81	0,26	0,14	0,94	0,8	0,98	0,99	0,92	0,62	0,058	0,073	0,97	0,97	0,5
4	0,45	0,31	0,27	0,56	0,069	0,37	0,75	0,8	0,79	0,66	0,37	0,57	**0,01**	0,069	0,66	0,54	0,35
5	0,32	0,61	0,68	0,39	0,15	0,47	0,63	0,82	0,41	0,086	0,28	0,33	**0,015**	0,061	0,65	0,64	0,42
6	0,65	0,14	0,19	0,44	**0,044**	0,28	0,28	0,45	0,12	0,27	0,57	0,18	**0,0063**	**0,029**	0,36	0,4	0,1
7	0,9	0,86	0,62	0,82	0,49	0,78	0,8	0,94	0,94	0,72	0,92	0,85	0,06	0,4	0,94	0,68	0,6
8	0,63	0,58	0,98	0,53	**0,022**	0,49	0,23	0,42	0,63	0,69	0,38	0,18	**0,0019**	0,1	0,084	0,15	0,1
9	0,66	0,88	0,72	0,35	0,25	0,71	0,55	0,2	0,58	0,52	0,54	**0,0036**	**0,01**	0,19	0,51	0,58	0,38
10	0,74	0,91	0,59	0,96	0,21	0,24	0,64	0,58	0,83	0,9	0,96	0,55	0,078	0,48	0,25	0,52	0,33
11	1	0,82	0,73	1	0,44	0,92	0,26	0,87	0,58	0,89	0,64	0,36	0,11	0,27	0,94	0,5	0,19
12	0,085	0,13	0,4	0,64	0,22	0,88	0,24	0,88	0,11	0,65	0,36	0,3	**0,03**	0,27	0,52	0,79	0,32
13	0,066	0,21	0,21	0,37	0,11	0,19	0,68	0,071	0,78	0,46	0,44	0,39	**0,008**	0,68	0,4	0,39	0,26
14	0,79	0,84	0,94	0,16	0,51	0,84	0,9	0,9	0,98	1	0,22	0,47	**0,083**	0,91	0,31	0,99	0,75
15	1	0,78	1	1	0,11	0,55	0,6	1	0,56	0,38	0,63	0,91	**0,0032**	0,82	0,74	0,21	0,13
16	0,75	0,053	0,95	0,81	0,02	0,11	0,58	0,69	0,58	0,32	0,55	0,55	**0,01**	0,61	0,47	0,65	0,15
17	0,69	0,24	0,057	0,29	0,08	0,12	**0,023**	0,4	0,21	0,53	0,27	0,22	**0,0065**	0,082	0,43	0,42	0,26
18	0,56	0,39	0,96	0,82	0,28	0,8	0,91	0,94	0,93	1	0,31	0,34	**0,0018**	0,51	0,54	0,91	0,63
19	0,93	0,95	0,96	0,89	**0,032**	0,82	0,87	0,93	0,9	0,97	0,76	0,91	**0,029**	0,8	0,77	0,57	0,78
20	0,95	0,45	0,53	0,45	0,058	0,45	0,65	0,83	0,31	0,66	0,6	0,59	**0,0034**	0,13	0,8	0,8	0,39
21	0,96	0,53	0,87	0,88	0,055	0,15	0,81	0,95	0,69	1	0,1	0,57	0,11	0,74	0,8	0,93	0,68
22	0,89	0,45	0,88	0,88	0,37	0,92	0,1	0,92	0,054	0,82	0,79	0,29	**0,023**	0,28	1	0,69	0,76
23	**0,032**	0,058	0,055	0,37	0,061	0,15	0,31	0,35	0,27	0,092	0,051	0,28	**0,0036**	0,11	0,11	0,076	0,098
24	0,82	0,45	0,15	0,92	0,15	0,61	0,53	0,53	0,73	0,7	0,89	0,8	**0,022**	0,17	0,79	0,78	0,091
25	0,87	0,65	0,81	0,1	0,31	0,53	0,55	0,72	0,27	0,11	0,57	0,5	**0,049**	0,23	0,77	0,67	0,35
26	0,93	0,83	0,95	0,92	0,35	0,53	0,72	0,65	0,48	0,49	0,098	0,31	**0,047**	0,17	0,87	0,48	0,18
27	0,9	0,31	0,69	0,054	0,27	0,73	0,27	0,48	0,82	0,92	0,29	0,29	**0,026**	**0,0033**	0,97	0,94	0,75
28	0,97	0,66	1	0,82	0,092	0,7	0,11	0,49	0,92	0,89	0,84	0,87	0,064	0,16	1	0,91	0,47
29	0,76	0,6	0,1	0,79	0,051	0,89	0,57	0,098	0,29	0,84	0,47	0,39	**0,025**	0,18	0,8	**0,025**	0,39
30	0,91	0,59	0,57	0,29	0,28	0,8	0,5	0,31	0,29	0,87	0,39	0,34	**0,012**	0,22	0,58	0,2	0,11
31	**0,029**	**0,0034**	0,11	**0,023**	**0,0036**	**0,022**	**0,049**	**0,047**	**0,026**	0,064	**0,025**	**0,012**	**0,002**	**0,0065**	**0,044**	**0,015**	**0,0025**
32	0,8	0,13	0,74	0,28	0,11	0,17	0,23	0,17	**0,0033**	0,16	0,18	0,22	**0,0065**	**0,039**	0,09	0,1	**0,027**
33	0,77	0,8	0,8	1	0,11	0,79	0,77	0,87	0,97	1	0,8	0,58	**0,044**	0,09	0,85	0,94	0,28
34	0,57	0,8	0,93	0,69	0,076	0,78	0,67	0,48	0,94	0,91	**0,025**	0,2	**0,015**	0,1	0,94	0,54	0,5
35	0,78	0,39	0,68	0,76	0,098	0,091	0,35	0,18	0,75	0,47	0,39	0,11	**0,0025**	**0,027**	0,28	0,5	0,14

TABLE 5. Results that are significant by either the chi-square goodness-of-fit test, or the algebraic method, or both.

Interaction	**Cancer**	**Hypothesis**	**p-value**	**Minimum expected value**	**p-value from DS-algorithm**
1-18	Bladder	FALSE	0,04	3,7	0,12
2-34	Bladder	FALSE	0,01	7,2	0,048
5-7	Bladder	FALSE	0,04	3,2	0,13
5-24	Bladder	FALSE	0,01	1,1	0,02
5-25	Bladder	FALSE	0,04	6,8	0,1
5-32	Bladder	FALSE	0,01	5,3	0,025
6-7	Bladder	FALSE	0,01	2,5	0,032
6-35	Bladder	FALSE	0,02	6,8	0,081
7-25	Bladder	FALSE	0,04	2,5	0,14
8-21	Bladder	FALSE	0,04	0,75	0,11
9-25	Bladder	FALSE	0	6,5	0,009
9-28	Bladder	FALSE	0,02	1,4	0,058
9-34	Bladder	FALSE	0,04	11	0,2
10-33	Bladder	FALSE	0,02	4,7	0,092
11-24	Bladder	impossible		0	0,03
11-25	Bladder	FALSE	0,02	0,36	0,039
11-33	Bladder	FALSE	0,02	0,36	0,075
11-34	Bladder	FALSE	0,04	0,36	0,18
12-29	Bladder	FALSE	0,03	1,4	0,06
14-29	Bladder	FALSE	0,01	0,74	0,014
14-34	Bladder	FALSE	0,03	1,9	0,1
15-33	Bladder	FALSE	0,02	2,9	0,088
17-21	Bladder	FALSE	0,03	0,37	0,074
17-33	Bladder	FALSE	0,04	5,1	0,13
18-34	Bladder	FALSE	0,04	5,6	0,15
20-34	Bladder	FALSE	0,01	1,5	0,049
21-34	Bladder	FALSE	0	1,9	0,0099
22-27	Bladder	FALSE	0,04	0,74	0,097
24-25	Bladder	FALSE	0,04	0,73	0,11
24-34	Bladder	FALSE	0,03	2,5	0,17
25-28	Bladder	FALSE	0,04	1,4	0,1
25-34	Bladder	FALSE	0,02	10	0,062
26-34	Bladder	FALSE	0,02	0,36	0,072
31-34	Bladder	impossible		0	0,038
33-34	Bladder	FALSE	0,01	13	0,081
34-34	Bladder	impossible		0	0,029
34-35	Bladder	FALSE	0,02	20	0,13

Interaction	Cancer	Hypothesis	p-value	Minimum expected value	p-value from D-S algorithm
1-24	Lung	FALSE	0,01	0,98	0,033
2-24	Lung	FALSE	0,01	0,66	0,026
3-24	Lung	impossible		0	0,0052
4-6	Lung	impossible		0	0,012
4-24	Lung	impossible		0	0,00077
4-25	Lung	impossible		0	0,034
5-10	Lung	FALSE	0,02	5,3	0,041
5-14	Lung	FALSE	0,01	0,67	0,014
5-22	Lung	FALSE	0,01	2,7	0,019
5-24	Lung	FALSE	0	0,66	0,0035
6-6	Lung	impossible		0	0,027
6-7	Lung	FALSE	0,04	0,33	0,066
6-8	Lung	FALSE	0,01	3,3	0,023
6-10	Lung	FALSE	0,01	2,3	0,039
6-12	Lung	FALSE	0,04	0,67	0,069
6-13	Lung	FALSE	0,02	1,7	0,048
6-14	Lung	FALSE	0	0,67	0,00059
6-15	Lung	FALSE	0,02	0,68	0,036
6-16	Lung	impossible		0	0,02
6-18	Lung	FALSE	0,01	3	0,018
6-21	Lung	impossible		0	0,047
6-22	Lung	FALSE	0,01	1	0,023
6-24	Lung	FALSE	0	0,66	0,002
6-27	Lung	FALSE	0,04	1	0,16
6-29	Lung	FALSE	0,04	1,7	0,1
6-33	Lung	FALSE	0,03	3,3	0,12
6-34	Lung	FALSE	0,02	4,3	0,089
6-35	Lung	FALSE	0,04	6,4	0,17
7-24	Lung	FALSE	0,01	0,98	0,021
8-22	Lung	FALSE	0,03	1,4	0,062
8-24	Lung	impossible		0	0,0085
9-24	Lung	FALSE	0,02	1,3	0,046
10-14	Lung	FALSE	0,02	0,67	0,054
10-22	Lung	FALSE	0,04	2	0,076
10-24	Lung	FALSE	0	0,66	0,00026
10-35	Lung	FALSE	0,04	9,5	0,18
11-24	Lung	impossible		0	0,011
12-24	Lung	impossible		0	0,028
13-22	Lung	FALSE	0,01	1	0,011
13-24	Lung	FALSE	0,02	0,34	0,05

Interaction	Cancer	Hypothesis	p-value	Minimum expected value	p-value from D-S algorithm
14-14	Lung	impossible		0	0,025
14-22	Lung	FALSE	0,02	0,67	0,041
14-24	Lung	impossible		0	0,0021
14-25	Lung	FALSE	0,04	0,34	0,11
14-26	Lung	FALSE	0,04	0,34	0,11
14-28	Lung	FALSE	0,01	0,33	0,02
14-33	Lung	FALSE	0,02	1,3	0,072
14-34	Lung	FALSE	0,04	2	0,12
14-35	Lung	FALSE	0,03	2	0,11
15-24	Lung	FALSE	0,04	0,35	0,13
16-24	Lung	impossible		0	0,0087
17-24	Lung	FALSE	0,01	1	0,022
17-28	Lung	FALSE	0,01	0,34	0,027
17-35	Lung	FALSE	0,04	6,9	0,15
19-24	Lung	impossible		0	0,0086
20-24	Lung	impossible		0	0,039
21-24	Lung	impossible		0	0,037
22-23	Lung	FALSE	0,03	0,68	0,054
22-24	Lung	FALSE	0,01	0,34	0,02
22-27	Lung	FALSE	0,01	0,68	0,023
22-28	Lung	FALSE	0,04	0,34	0,067
22-29	Lung	FALSE	0	1	0,00035
23-24	Lung	FALSE	0,01	0,99	0,04
24-24	Lung	impossible		0	0,002
24-25	Lung	FALSE	0	0,66	0,0084
24-26	Lung	impossible		0	0,043
24-27	Lung	FALSE	0,01	0,66	0,0092
24-28	Lung	impossible		0	0,014
24-29	Lung	FALSE	0,02	0,65	0,028
24-30	Lung	impossible		0	0,0096
24-30	Lung	impossible		0	0,0077
24-32	Lung	FALSE	0,01	0,32	0,016
24-33	Lung	FALSE	0	0,65	0,013
24-34	Lung	FALSE	0	1,9	0,021
24-35	Lung	FALSE	0	1,7	0,016
29-32	Lung	FALSE	0,01	3,3	0,036
34-35	Lung	FALSE	0,02	23	0,12

Interaction	Cancer	Hypothesis	p-value	Minimum expected value	p-value from D-S algorithm
1-17	Leukaemia	FALSE	0,04	2,2	0,082
2-27	Leukaemia	FALSE	0	2,6	0,0072
2-21	Leukaemia	impossible		0	0,0095
2-23	Leukaemia	FALSE	0,04	2,8	0,11
2-31	Leukaemia	FALSE	0,04	0,37	0,052
2-32	Leukaemia	FALSE	0,03	5	0,07
4-6	Leukaemia	impossible		0	0,048
4-31	Leukaemia	impossible		0	0,01
5-8	Leukaemia	FALSE	0,04	5,3	0,11
5-16	Leukaemia	impossible		0	0,021
5-17	Leukaemia	FALSE	0,01	5,1	0,026
5-28	Leukaemia	FALSE	0,04	0,35	0,086
5-31	Leukaemia	impossible		0	0,015
5-32	Leukaemia	FALSE	0,02	5	0,061
6-8	Leukaemia	FALSE	0,01	3,2	0,029
6-9	Leukaemia	FALSE	0	5,7	0,016
6-11	Leukaemia	FALSE	0,01	0,36	0,016
6-15	Leukaemia	FALSE	0,01	3,3	0,017
6-23	Leukaemia	FALSE	0,01	3,9	0,044
6-31	Leukaemia	FALSE	0,01	0,37	0,0063
6-32	Leukaemia	FALSE	0,01	3,6	0,029
6-35	Leukaemia	FALSE	0,03	7,8	0,1
8-23	Leukaemia	FALSE	0	6,4	0,022
8-31	Leukaemia	FALSE	0	0,37	0,0019
8-32	Leukaemia	FALSE	0,04	2,9	0,1
8-33	Leukaemia	FALSE	0,02	6,5	0,084
8-34	Leukaemia	FALSE	0,04	13	0,15
8-35	Leukaemia	FALSE	0,02	11	0,1
9-12	Leukaemia	FALSE	0,01	1,4	0,011
9-17	Leukaemia	FALSE	0	5,1	0,011
9-30	Leukaemia	FALSE	0	0,36	0,0036
9-31	Leukaemia	impossible		0	0,01
12-31	Leukaemia	impossible		0	0,03
13-15	Leukaemia	FALSE	0,03	1,1	0,057
13-23	Leukaemia	FALSE	0,04	2,1	0,11
13-26	Leukaemia	FALSE	0,03	1,4	0,071
13-31	Leukaemia	impossible		0	0,008
15-17	Leukaemia	FALSE	0,04	2,6	0,11
15-31	Leukaemia	impossible		0	0,0032

Interaction	**Cancer**	**Hypothesis**	**p-value**	**Minimum expected value**	**p-value from D-S algorithm**
15-35	Leukaemia	FALSE	0,03	6,1	0,13
16-23	Leukaemia	impossible		0	0,02
16-31	Leukaemia	impossible		0	0,01
17-21	Leukaemia	FALSE	0,03	1,1	0,057
17-23	Leukaemia	FALSE	0,03	5,4	0,08
17-24	Leukaemia	FALSE	0,04	0,36	0,12
17-25	Leukaemia	FALSE	0,01	5,1	0,023
17-31	Leukaemia	impossible		0	0,0065
17-32	Leukaemia	FALSE	0,04	4,7	0,082
18-31	Leukaemia	impossible		0	0,0018
19-23	Leukaemia	impossible		0	0,032
19-31	Leukaemia	impossible		0	0,029
20-23	Leukaemia	FALSE	0,03	0,72	0,058
20-31	Leukaemia	impossible		0	0,0034
21-23	Leukaemia	FALSE	0,02	1,1	0,055
21-29	Leukaemia	FALSE	0,04	0,36	0,1
22-31	Leukaemia	impossible		0	0,023
23-28	Leukaemia	FALSE	0,04	0,7	0,092
23-29	Leukaemia	FALSE	0,03	5,7	0,051
23-31	Leukaemia	impossible		0	0,0036
23-33	Leukaemia	FALSE	0,04	9	0,11
23-34	Leukaemia	FALSE	0,02	16	0,076
23-35	Leukaemia	FALSE	0,03	12	0,098
24-31	Leukaemia	impossible		0	0,022
24-35	Leukaemia	FALSE	0,02	0,69	0,091
25-31	Leukaemia	impossible		0	0,049
26-29	Leukaemia	FALSE	0,04	1,1	0,098
26-31	Leukaemia	impossible		0	0,047
27-31	Leukaemia	impossible		0	0,026
27-32	Leukaemia	FALSE	0	3,2	0,0033
29-31	Leukaemia	impossible		0	0,025
29-34	Leukaemia	FALSE	0	7,9	0,025
30-31	Leukaemia	impossible		0	0,012
31-31	Leukaemia	impossible		0	0,002
31-32	Leukaemia	impossible		0	0,0065
31-33	Leukaemia	FALSE	0,02	0,37	0,044
31-34	Leukaemia	FALSE	0,01	0,36	0,015
31-35	Leukaemia	impossible		0	0,0025
32-32	Leukaemia	impossible		0	0,039
32-33	Leukaemia	FALSE	0,02	4,3	0,09
32-34	Leukaemia	FALSE	0,03	12	0,1
32-35	Leukaemia	FALSE	0	11	0,027

Acknowledgements. This work was made possible by the MeRiMa group of the Department of Mathematics of the University of Turin. We used data from the GenAir study. GenAir collaborator were: M. Manuguerra, G. Matullo, F. Veglia, H. Autrup, A.M. Dunning, S. Garte, E. Gormally, C. Malaveille, S. Guarrera, S. Polidoro, F. Saletta, M. Peluso, L. Airoldi, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, D. Trichopoulos, A. Kalandidi, D. Palli, V. Krogh, R. Tumino, S. Panico, H.B. Bueno-De-Mesquita, P.H. Peeters, E. Lund, G. Pera, C. Martinez, P. Amiano, A. Barricarte, M.J. Tormo, J.R. Quiros, G. Berglund, L. Janzon, B. Jarvholm, N.E. Day, N.E. Allen, R. Saracci, R. Kaaks, and P. Ferrari.
The fourth and the last authors were supported by the PRIN "Geometria delle varietà algebriche e dei loro spazi

di moduli", cofinanced by MIUR (Italy) (cofin 2008).
The authors declare they have no actual or potential competing financial interests.

References

[1] A. Agresti, *Exact inference for categorical data: Recent advances and continuing controversies*, Statist. Med. 20 (2001), 2709–2722.
[2] A. Agresti, Categorical data analysis, Wiley, 2002.
[3] H. Aurtrup, *Genetic polymorphisms in human xenobiotica metabolizing enzymes as susceptibility factors in toxic response*, Mutat Res 464 (2000), 65–76.
[4] N. Beerenwinkel, L. Pachter, B. Sturmfels, S.F. Elena, R.E. Lenski, *Analysis of epistatic interactions and fitness landscapes using a new geometric approach.*, BMC Evol Biol. 13 (2007), 7:60.
[5] S.P. Cleary, M. Cotterchio, E. Shi, S. Gallinger, P. Harper, *Cigarette smoking, genetic variants in carcinogen-metabolizing enzymes, and colorectal cancer risk*, Am. J. Epidemiol. 172 (2010), no. 9, 1000–1014.
[6] H.J. Cordell, *Detecting gene-gene interactions that underlie human diseases*, Nat Rev Genet, 10 (2009), 392–404.
[7] D. Cox, J. Little, D. O'Shea, Ideals, varieties, and algorithms, Undergraduate Texts in Mathematics, vol. 60, Springer-Verlag, New York, 1992.
[8] A.C. Davison, D.V. Hinkley, Bootstrap methods and their applications, Cambridge University Press, Cambridge, 1997.
[9] P. Diaconis, B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Ann. Statist., 26 (1998), 363–397.
[10] M. Drton, S. Sullivant, *Algebraic statistical model*, Statist. Sinica., 17 (2007), 1273–1297.
[11] F. Dudbridge, A. Gusnanto, B.P.C. Koeleman, *Detecting multiple associations in genome-wide studies*, Human Genomics, 2 (2006), 310–317.
[12] F. Dudbridge, B.P.C. Koeleman, *Efficient computation of signifcance levels for multiple associations in large studies of correlated data, including genomewide association studies*, Am. J. Hum. Genet, 75 (2004), 424–435.
[13] E.S. Edgington, Randomization tests (3rd ed.), Marcel Dekker, New York, 1995.
[14] B. Efron, *The jackknife, the bootstrap and other resampling plans*, Society of Industrial and Applied Mathematics CBMS-NFS Monographs, vol. 38, Capital City Press, Philadelphia, 1982.
[15] L Fan, J.O. Fuss, Q.J. Cheng, A.S. Arvai, M. Hammel, V.A. Roberts, P.K. Cooper, J.A. Tainer, *XPD helicase structures and activities: insights into the cancer and aging phenotypes from xpd mutations.*, Cell, 133 (2008), 789–800.
[16] C. Fassino, M.L. Torrente, *Simple approximate varieties for sets of empirical points*, Submitted. Available at http://arxiv.org/abs/1008.0274
[17] I.O. Filiz, X. Guo, J. Morton, B. Sturmfels, *Graphical models for correlated defaults*, Available at http://arxiv.org/pdf/0809.1393v1.pdf, 2008.
[18] R.A. Fisher, The design of experiments, Oliver and Boyd, Edinburgh, 1935.
[19] W. Fulton, Introduction to toric varieties, Princeton University Press, 1993.
[20] P. Good, Resampling methods: A practical guide to data analysis (3rd edition), Birchäuser, Boston, 2006.
[21] H. Gorji, N Shahbazi, P. Habibollahi, S.M. Tavangar, A. Firooz, M.H. Ghahremani, *The glutathione-S-transferase P1 polymorphisms correlates with changes in expression of TP53 tumor suppressor in cutaneous basal cell carcinoma*, Dermatol Sci 56 (2009), 208–10.
[22] L.W. Hahn, M.D. Ritchie, J.H. Moore, *Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions*, Bioinformatics, 19 (2003), 376–382.
[23] I. Hallgrimsdottir, B Sturmfels, *Resultants in genetic linkage analysis*, Journal of Symbolic Computation, 41 (2006), 125–137.
[24] D.Y. Lin, *An efficient monte carlo approach to assessing statistical significance in genomic studies*, Bioinformatics, 21 (2005), 781–787.
[25] H.W. Lo, L. Stephenson, X. Cao, M. Milas, R. Pollock, F. Ali-Osman, *Identification and functional characterization of the human glutathione S-transferaseP1 gene as a novel transcriptional target of the p53 tumor suppressor gene.*, Mol Cancer Res, 6 (2008), 843–50.
[26] A.S. Malaspinas, C. Uhler, *Detecting epistases via markov bases*, Journal of Algebraic Statistics, 2 (2011), no. 1, 36–53.
[27] M. Manuguerra, G. Matullo, F. Veglia, H. Autrup, A.M. Dunning, S. Garte, E. Gormally, C. Malaveille, S. Guarrera, S. Polidoro, F. Saletta, M. Peluso, L. Airoldi, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, D. Trichopoulos, A. Kalandidi, D. Palli, V. Krogh, R. Tumino, S. Panico, H.B. Bueno-De Mesquita, P.H. Peeters, E. Lund, G. Pera, C. Martinez, P. Amiano, A. Barricarte, M.J. Tormo, J.R. Quiros, G. Berglund, L. Janzon, B. Jarvholm, N.E. Day, N.E. Allen, R. Saracci, R. Kaaks, P. Ferrari, E. Riboli, P. Vineis, *Multi-factor dimensionality reduction applied to a large prospective investigation on gene-gene and gene-environment interactions*, Carcinogenesis, 28(2) (2007), 414–22.
[28] T. Martone, P. Vineis, C. Malaveille, B. Terracini, *Impact of polymorphisms in xeno(endo)biotic metabolism on pattern and frequency of p53 mutations in bladder cancer.*, Mutat Res, 462 (2000), 303–9.
[29] G. Matullo, A.M. Dunning, S. Guarrera, C. Baynes, S. Polidoro, S. Garte, H. Autrup, C. Malaveille, M. Peluso, L. Airoldi, F. Veglia, E. Gormally, G. Hoek, M. Krzyzanowski, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, A. Trichopoulou, D. Palli, V. Krogh, R. Tumino, S. Panico, H.B. Bueno-De Mesquita, P.H. Peeters, E. Lund, G. Pera, C. Martinez, M. Dorronsoro, A. Barricarte, M.J. Tormo, J.R. Quiros, N.E. Day, T.J. Key, R. Saracci, R. Kaaks, E. Riboli, P. Vineis, *DNA repair polymorphisms and cancer risk in non-smokers in a cohort study*, Carcinogenesis, 27(5) (2006), 997–1007.

[30] Y. Meng, Q. Ma, Y. Yu, J. Farrell, L.A. Farrer, M.A. Wilcox, *Multifactor-dimensionality reduction versus family-based association tests in detecting susceptibility loci in discordant sib-pair studies.*, BMC Genet, 30(6) (2005), S146.

[31] J. Molitor, M. Papathomas, M Jerrett, and S. Richardson, *Bayesian profile regression with an application to the national survey of children's health.*, Biostatistics, 11 (2010), 484–498.

[32] D.S. Moore, G. McCabe, W. Duckworth, S. Sclove, Chapter 18:bootstrap methods and permutation tests, The Practice of Business Statistics, W.H. Freeman, New York, 2003.

[33] L. Pachter, B. Sturmfels, *Parametric inference for biological sequence analysis*, Proc Natl Acad Sci U S A, 101 (2004), 16138–43.

[34] L. Pachter, B. Sturmfels, *Tropical geometry of statistical models*, Proc Natl Acad Sci U S A, 101 (2004), 16132–7.

[35] M. Papathomas, J. Molitor, S. Richardson, E. Riboli, P. Vineis, *Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers*, Environ. Health Perspect, 119 (2011), 84–91.

[36] L. Patchter, B. Sturmfels, *Algebraic statistics for computational biology*, Cambridge University Press, 2005.

[37] M. Peluso, P. Hainaut, L. Airoldi, H. Autrup, A. Dunning, S. Garte, E. Gormally, C. Malaveille, G. Matullo, A. Munnia, E. Riboli, P. Vineis, *Methodology of laboratory measurements in prospective studies on gene-environment interactions: the experience of GenAir*, Mutat Res, 574 (2005), 92–104.

[38] G. Pistone, E. Riccomagno, and H.P. Wynn, Algebraic statistics, Chapman and Hall/CRC, Boca Raton, 2001.

[39] F. Rapallo, *Algebraic Markov bases and MCMC for two-way contingency tables*, Scandinavian Journal of Statistics, 30 (2003), 385–397.

[40] F. Rapallo, *Algebraic exact inference for rater agreement models*, Statistical Methods & Applications, 14 (2005), 45–66.

[41] E. Riboli, *The european prospective investigation into cancer and nutrition (EPIC): plans and progress.*, J. Nutr., 131 (2001), no. 1, 170–175.

[42] T.K. Rice, N.J. Schork, D.C. Rao, *Methods for handling multiple testing*, Advances in Genetics, 60 (2008), 293–308.

[43] M.D. Ritchie, L.W. Hahn, N Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, *Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer*, Am. J. Hum. Genet., 69 (2001), no. 1, 138–47.

[44] J.L. Simon, Resampling: The new statistics (2nd edition), http://bcs.whfreeman.com/pbs/, 1997.

[45] B. Sturmfels, Gröbner bases and convex polytopes, American Mathematical Society, 1996.

[46] B. Sturmfels, Solving systems of polynomial equations, American Mathematical Society, 2002.

[47] B. Sturmfels, Algebra and geometry of statistical models, Tech. report, John von Neumann Lectures, TU München, 2003.

[48] B. Sturmfels, S. Sullivant, *Toric ideals of phylogenetic invariants*, J Comput Biol, 12 (2005), 204–228.

[49] P. Vineis, L. Airoldi, F. Veglia, L. Olgiati, R. Pastorelli, H. Autrup, A. Dunning, S. Garte, E. Gormally, P. Hainaut, C. Malaveille, G. Matullo, M. Peluso, K. Overvad, A. Tjonneland, F. Clavel-Chapelon, H. Boeing, V. Krogh, D. Palli, S. Panico, R. Tumino, B. Bueno-De Mesquita, P. Peeters, G. Berglund, G. Hallmans, R. Saracci, E. Riboli, *Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study.*, BMJ 330 (2005), 277.

[50] S. Wang, W. Xiong, W. Ma, S. Chanock, W. Jedrychowski, R. Wu, F.P. Perera, *Gene-environment interactions on growth trajectories*, Genetic Epidemiology (2012), doi: 10.1002/gepi.21613.

[51] R.D. Wood, *Mammalian nucleotide excision repair proteins and interstrand crosslink repair*, Environ Mol Mutagen, 51 (2010), 520–6.

[52] Y. Zhang, J.S. Liu, *Bayesian inference of epistatic interactions in case-control studies.*, Nature Genet, 39 (2007), 1167–1173.

[53] Y. Zhang, L.H. Rohde, H. Wu, *Involvement of nucleotide excision and mismatch repair mechanisms in double strand break repair*, Curr Genomics, 10 (2009), 250–8.