View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

Coll. Antropol. **30** (2006) 2: 383–385
Original scientific paper

Surnames as Markers of Pathologies – Two Statistical Techniques and Their Applications

Marco B. L. Rocchi

Institute of Biomathematics, Faculty of Pharmacy, University of Urbino »Carlo Bo«, Urbino, Italy

A B S T R A C T

The objective of this research is to propose and to validate two different statistical techniques to test the hypothesis of an association between surnames and pathologies, in a population participating in a screening procedure for a given pathology. We propose two statistical methods: a first technique is based on the rarefaction method, and second one is based on the principle of resampling, and it can be considered a special case of a randomisation test. Both the techniques are applied to a data set of babies screened for congenital hypothyroidism (CH), and they gave similar results. The large overlapping of the results seems to suggest a substantial validity of the proposed techniques.

***Key words**: screening, surnames, congenital hypothyroidism*

Introduction

In societies where surnames pass from father to children, the Y chromosome is passed down in the same way that a surname.

Therefore, we can consider surnames as genetic character linked to the Y chromosome[1]. The difference lies in the fact that although all of the offspring get the surname, only half of them – the sons – get the Y chromosome.

Hence, surnames have been used effectively as markers of biodiversity among different populations, rather than other biological markers that are more complicated both to obtain and to study, such as hypervariable region of mtDNA[2]. There are already several applications for surnames in anthropology and biodemography, while the possible role of surnames in genetic analysis has only recently been discussed by Jobling[3].

A detailed review regarding the use of surnames in epidemiological research has been proposed by Cook et al.[4]. They considered two different applications for surnames, both as indicators of ethnic or national origin, and as part of the identifying information needed for record linkage or follow-up studies.

More recently, several authors have used surnames as markers for pathologies[5–10], but these authors based their works on different assumptions: (a) surnames are useful tools for the identification of ethnicity in multiethnic populations. In this case, surnames provide an easy and effective means of subdivision for epidemiological purposes[7]; (b) the role of surnames as pathology markers may be explained as a consequence of the founder effect, especially in the case of geographical (and also genetic) isolates; (c) from a broader standpoint, surnames can be considered »alleles« of a genetic locus that is associated with another genetic locus coding for a pathology. As an example of the last consideration, a surname analysis of over 40,000 Wisconsin cancer mortalities over the period 1979–1985 was performed by Cleek[8] to demonstrate the genetic component associated with various major cancers (i.e. male and female leukaemia, and male lung cancer).

In this paper, we propose two different statistical techniques to test the hypothesis that some surnames are linked to certain pathologies.

Obviously, this approach doesn't take into account the factor that probably is the most important in case of genetic disorders of recessive inheritance: the risk of the disease is high in familiar groups with high degree of consanguinity, where individuals are sharing the same surnames. Nevertheless, the approaches presented in this study should actually be interpreted as an exploratory tool to detect the association between surnames and pathologies, without taking into consideration the pathological model of the disease.

Moreover, we only consider the case of pathologies that are screened for.

Received for publication April 19, 2004

Materials and Methods

The first statistical technique that we propose is based on the rarefaction method, developed by Sanders[10], and corrected and modified by Simberloff[11]. This technique is usually applied in biodiversity analysis.

The second statistical technique is an adjustment of a resampling method (i.e. a randomisation test).

The rarefaction method

This technique is based on the assumption that we know the surname distribution of the population participating in the screening procedure. More specifically, we know:

N – number of screened people;
S – number of surnames in the screened population;
n – number of subjects that test positive in the screening ($n<N$);
x – number of surnames in the set of positive subjects ($x \leq n$);
N_i – number of subjects in the population, with the i-th surname ($i=1,2,\ldots,S$), so that: $\sum_{i=1}^{S} N_i = N$.

Both the expected value $E(x)$ and the variance $Var(x)$ of x can be obtained using the rarefaction method:

$$E(x) = \sum_{i=1}^{S} \left[1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right]$$

$$Var(x) = \binom{N}{n}^{-1} \left\{ \sum_{i=1}^{S} \binom{N-N_i}{n} \left[1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right] + \right.$$

$$\left. +2 \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} \left[\binom{N-N_i-N_j}{n} - \frac{\binom{N-N_i}{n}\binom{N-N_j}{n}}{\binom{N}{n}} \right] \right\}$$

Note that $E(x)$ can be interpreted as the number of surnames expected in a sample of size n, randomly extracted from the known distribution of surnames.

Once the $E(x)$ and $Var(x)$ are known, we can easily perform a z-test; if the observed value of x is significantly smaller than the expected value of x, then we can reject the null hypothesis.

The statistics $z = \dfrac{x - E(x)}{\sqrt{Var(x)}}$ is approximately distributed as a standard normal distribution under the null hypothesis. The test can be performed as a one-tailed test, because it is reasonable to fix the alternative hypothesis unidirectionally.

The decision rule (at a 5% significance level) is: if $z<-1.645$ we can reject the null hypothesis and we can conclude that there is an association between surnames and the given pathology.

The randomisation method

Once the surname distribution in the population of subjects participating in the screening is known, we can randomly extract a high number of samples (at least 1,000) having the same size as the subpopulation that test positive in the screening (this procedure is named resampling).

For each sample the number of observed surnames is counted. We can then arrange all the samples according to the number of recorded surnames. Among these samples, we record the position occupied by the positive subpopulation. The decision rule is then applied, i.e.: if the positive subpopulation occupies a position within the 5-th percentile (one-tailed test) we can reject the null hypothesis and conclude that there is an association between surnames and the pathology. The detailed algorithm for this randomisation test is the following:

- step 1: compute the number of surnames in the positive subpopulation (denote this number with the symbol x^* and the size of the positive subpopulation with n);
- step 2: randomly extract k samples (with $k=1{,}000$) of size n from the population (of size $N>n$);
- step 3: for each of the k samples, compute the number of observed surnames (denote this number with the symbol x_k);
- step 4: arrange the $(k+1)$ results (i.e., the x^* and all the x_k) in an increasing order, and denote the ordered x_k with the symbol $x_{(k)}$);
- step 5: note the position r of x^*;
- step 6: if $r<k\times0.05$ we can reject the null hypothesis;
- step 7: compute the p-value as $p=r/k$.

Results

Data set

We have considered the complete database of neonatal screening for Congenital Hypothyroidism (CH) in the Marche region (Italy) in the 16-year period 1981–1996.

In this period, $N=186{,}690$ newborns participated in the screening and $n=92$ tested positive.

Results of rarefaction method

We observed $x=88$ different surnames in the set of $n=92$ babies testing positive for CH.

The expected number of surnames under the null hypothesis (i.e., no association between surname and pathology) was $E(x)=89.9$, and the variance of x was $Var(x)=1.83$.

The test statistics z results $z = \dfrac{88-89.9}{\sqrt{1.83}} = -1.404$.

According to the decision rule, we cannot reject the null hypothesis, and the observed value of probability

(p-value) resulting from the statistical table of the normal gaussian distribution is $p=0.0808$.

Results of randomisation method

We denoted with $x^*=88$ the number of different surnames out of $n=92$ babies testing positive for CH.

We performed $k=10{,}000$ resampling and for each sample we computed the number of surnames (x_k). We then rearranged the samples in increasing order ($x_{(k)}$). The position of x^* among the $x_{(k)}$ values coming from the resampling procedure was $r=714$.

Since 714 is greater than 500 ($=10{,}000\times0.05$), we cannot reject the null hypothesis, and the observed value of probability (p-value) was: $p=0.0714$.

Discussion and Conclusion

In this paper we proposed and analysed two different statistical techniques able to detect an association between surnames and pathologies, in the particular case of diseases for which a screening program is developed.

The first technique is based on the rarefaction method, while the second one can be considered a special case of the randomisation test.

When these techniques were applied to a data set of babies participating in a regional screening for congenital hypothyroidism (CH) the results appeared largely overlapping. Hence, the proposed methods appear to be substantially valid.

However, it should be emphasized that, in order to apply these techniques, the knowledge of the surname distribution in the screened population is necessary.

Acknowledgements

The author wishes to thank the anonymous referee for his helpful suggestions in improving the quality of the paper.

REFERENCES

1. CHEN, K. H., L. L. CAVALLI-SFORZA, Hum. Biol., 55 (1983) 367. — 2. STENICO, M., L. NIGRO, G. BERTORELLE, F. CALAFELL, M. CAPITANIO, C. CORRAIN, G. BARBUJANI, Am. J. Hum. Genet., 59 (1996) 1363. — 3. JOBLING, M. A., Trends Genet., 17 (2001) 353. — 4. COOK, D., D. HEWITT, J. MILNER, Am. J. Epidemiol., 95 (1972) 38. — 5. RUDAN, I., G. N. RANZANI, M. STRNAD, A. VORKO-JOVIC, V. JOHN, J. UNUSIC, D. IVANKOVIC, R. STEVANOVIC, S. VULETIC, P. RUDAN, Coll. Antropol., 23 (1999) 557. — 6. DE SILVESTRI, A., C. R. GUGLIELMINO, Hum. Biol., 72 (2000) 573. — 7. TORRINGTON, M., P. A. BRINK, S. Afr. Med. J., 77 (1990) 289. — 8. CLEEK, R. K., Hum. Biol., 61 (1989) 195. — 9. GARZA-CHAPA, R., M. DE LOS ANGELES ROJAS-ALVARADO, R. M. CERDA-FLORES, Am. J. Hum. Biol., 12 (2000) 721. — 10. RUDAN, I., Hum. Biol., 73 (2001) 871. — 11. SANDERS, H. L., Am. Nat., 102 (1968) 243. — 12. SIMBERLOFF, D. S., Am. Nat., 106 (1972) 414.

M. B. L. Rocchi

Institute of Biomathematics, Faculty of Pharmacy, University of Urbino »Carlo Bo«, Loc. Crocicchia, 61029 Urbino, Italy
e-mail: m.rocchi@uniurb.it

PREZIMENA KAO BILJEZI BOLESTI – DVIJE STATISTIČKE METODE I NJIHOVA PRIMJENA

SAŽETAK

Cilj ovog istraživanja je predložiti i procijeniti dvije različite statističke metode za testiranje hipoteze povezanosti prezimena i bolesti u populaciji testiranoj za određenu bolest. Mi predlažemo dvije statističke metode: prva metoda se temelji na rarefakcijskoj metodi a druga na ponovljenom odabiranju uzoraka i može se smatrati posebnim slučajem testa nasumičnog odabira. Obje metode primijenjene su na podacima o dojenčadi testiranim na urođeni hipotireoizam (CH) i dale su slične rezultate. Velika podudarnost rezultata upućuje na značajnu preciznost predloženih metoda.