# The cultural heritage shapes the pattern of tumour profiles in Europe: a correlation study

Romualdo Benigni, Rosa Giaimo, Domenica Matranga, Alessandro Giuliani

**Abstract**

***Study objective*—This study investigates the spatial pattern of tumours in Europe to check the feasibility of a large scale ecological epidemiology approach to cancer in Europe.**

***Setting*—The tumour types relative frequencies and cancer incidence (for men and women) reported in the European cancer registries were investigated by exploratory data analysis techniques. Socioeconomical descriptors of the female condition were considered as well.**

***Main results*—The classification of the European regional areas covered by the cancer registries followed almost exactly the boundaries set by the long and intermingled European history in terms of life styles and cultural heritage. This result supports the notion of a predominant role of environmental factors in cancer induction. Further support to the above result was given by the finding of a correlation between differential male-female cancer incidence, and socioeconomic descriptors of the female condition.**

***Conclusions*—From a methodological point of view, the consistency of these results pointed to the feasibility of an ecological approach to tumour epidemiology.**

(*J Epidemiol Community Health* 2000;**54**:262–268)

A large literature indicates that the multiplicity of cancer causes dramatically lowers the power of classic epidemiological studies in pointing out cancer risk factors,[1 2] thus giving rise to attempts to find new methodological tools, which can be classified into the large families of molecular epidemiology on the one hand, and ecological epidemiology on the other hand. This work is an ecological analysis of a very specific case (tumour induction and spectra in Europe); it is also aimed at contributing to the investigation on the practicality of such methods. In fact, the presence of overwhelming causal associations (like lung cancer and smoking habits), of exceedingly frequent tumour types (like lung and prostate for men and breast for women) and the huge spatial variability in average tumour incidence (around 50% for the areas considered in our study) seem to be, in principle, difficult obstacles for ecological studies on this subject.

Moreover, Europe has had a long and intermingled history in which, for more than 3000 years, populations strongly interacted with each other in all possible ways (war, commercial and cultural exchanges, massive migrations, etc).[3–5] These interactions not only simply followed the historical events but also shaped the cultural heritage of Europeans by generating a strong general commonality and equally strong regional differences.[4 5] The cultural differences among European populations, involving almost all aspects of everyday life, from food to the use of leisure time, are still present today (even if less evident than in the past decades) and, together with physical constraints such as climate, altitude, relative distance from the sea, shape the European environment. European history permits the sketching of some indistinct boundaries partitioning Europe into macro-areas with relatively homogeneous cultural heritage. These boundaries, given the high level of gene flux among Europeans through the centuries,[3] are correlated but not coincident with the areas delineated by population genetics[3 6 7]

Given these premises, and as environmental factors are recognised as predominant in the induction of cancer,[1] the general question we want to answer is: is it possible to exploit the above differences related to cultural heritage for modelling the geographical patterns of tumour incidence in Europe and recognising more clearly causal factors of cancer? The importance of this issue is related to the need for an efficient, epidemiologically-based cancer surveillance.

## Methods

### CANCER DATA

Table 1 lists the European geographical areas studied in this paper. These are the areas for which cancer registries, meeting the reliability criteria set up by the International Agency for Research on Cancer, are available. The areas analysed are those present in both 1985–1988 and 1988–1992 International Agency for Research on Cancer compilations.[8 9] This selection was dictated by the need to check for the consistency between the two compilations of data. This consistency was actually demonstrated in a previous work.[10] Each area was defined by 85 variables corresponding to the standardised proportion of the incidence of 41 male and 44 female types of tumours (see names in table 3). The proportion is normalised for the total incidence of tumours in each area. The data analysed in this work were retrieved from the International Agency for Research on Cancer compilation relative to the 1988–1992 period.[9]

The above data have a number of advantages for this analysis. The macro scale of the regional areas, being on average over millions of people, is not influenced by the interindividual variability as in the classic

**Istituto Superiore di Sanita', Toxicology and Ecotoxicology Laboratory,Viale Regina Elena 299, 00161 Rome, Italy**
R Benigni
A Giuliani

**Department of Statistics, Faculty of Economics, University of Palermo, Italy**
R Giaimo

**Italian Institute of Statistics ISTAT, Palermo Branch**
D Matranga

Correspondence to:
Dr Benigni

Accepted for publication
12 August 1999

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.

*Table 1 European regions relative to the cancer registries analysed*

| *European areas* | *Codes* |
|---|---|
| Belarus | Belar |
| Czech Republic, Bohemia and Moravia | Boem |
| Denmark | Denm |
| Finland | Finl |
| France, Bas Rhin | Frabr |
| France, Calvados | Fracal |
| France, Doubs | Fradou |
| France, Isère | Fraise |
| France, Somme | Frasom |
| France, Tarn | Fratar |
| Germany, Berlin | Gerest |
| Germany, Saarland | Gersar |
| Iceland | Iceland |
| Ireland, Cork | Eire |
| Italy, Florence | Itafir |
| Italy, Genoa | Itagen |
| Italy, Latina | Italat |
| Italy, Varese | Itavar |
| Italy, Parma | Itapar |
| Italy, Ragusa | Itarag |
| Italy, Romagna | Itafor |
| Italy, Turin | Itator |
| Italy, Trieste | Itatrie |
| Norway | Norw |
| Poland, Lower Silesia | Polsle |
| Poland, Warsaw | Polwar |
| Slovakia | Slovak |
| Slovenia | Sloven |
| Spain, Tarragon | Spatar |
| Spain, Granada | Spagra |
| Spain, Murcia | Spamur |
| Spain, Navarra | Spanav |
| Spain, Zaragoza | Spazar |
| Sweden | Swed |
| UK, England and Wales | Uktot |
| UK, Birmingham | Ukbir |
| UK, Mersey | Ukmer |
| UK, North Western | Ukmanc |
| UK, Oxford | Ukox |
| UK, South Thames | Uksth |
| UK, South Western | Ukswr |
| UK, Yorkshire | Ukyl |
| UK, Scotland | Scotot |
| UK, West Scotland | Scowes |

epidemiological studies, thus being suitable for highlighting large scale trends that interindividual variability often masks. On the other hand, the fact that we relied on the variability between normalised tumour spectra automatically ruled out the problems related to the presence of exceedingly frequent cancers, to the global tumour incidence variability between areas (probably linked to peculiar exposition situations) and to the presence of very strong (and thus uniformly distributed) causal associations.

For the purpose of this analysis, the 85 variables that represent the tumour types (tumour spectra) were summarised into five Principal Components for the female population (PCF1 to PCF5), and five Principal Components for the male population (PCM1 to PCM5) (table 3). A short presentation of the Principal Component Analysis is given below.

In addition to the profile variables, the normalised difference between male and female global tumour incidence (DELTAN) was used to characterise the studied areas (table 2) (DELTAN= (PM-PF)/PM, with PM representing the whole incidence of tumours in men and PF the whole incidence of tumours in women normalised per 10000 inhabitants).

### SOCIOECONOMIC VARIABLES

Socioeconomic data relative to the female condition were available for 37 European countries. These included 16 countries for which we had the pathology data analysed here, plus Albania, Austria, Belgium, Bosnia, Bulgaria, Croatia, Estonia, Greece, Hungary, Latvia, Lithuania, Luxembourg, Macedonia, Malta, Netherlands, Portugal, Romania, Russia, Switzerland, Ukraine, Yugoslavia. The demographic variables are listed in table 5. These 15 variables gave rise to four Principal Components (DEM1 to DEM4), collectively explaining 81% of variance; Component 1 (DEM1) by itself explained 46% of total variability. DEM1-DEM4 were used to investigate the correlation between the demography and pathology description of the European countries.

*Table 2 Sorted DELTAN values. These data show an extremely high variability of differential cancer incidence, ranging from the quasi equivalence between sexes in Denmark to the huge difference in Calais*

| *Area* | *DELTAN* | *Area* | *DELTAN* |
|---|---|---|---|
| Denm | 0.067 | Itafir | 0.331 |
| Ukox | 0.073 | Gersar | 0.339 |
| Sweden | 0.075 | Boem | 0.347 |
| Ukmer | 0.093 | Italat | 0.355 |
| Ukswr | 0.093 | Itator | 0.364 |
| Iceland | 0.113 | Sloven | 0.367 |
| Uksth | 0.114 | Polsle | 0.370 |
| Eire | 0.148 | Spanav | 0.374 |
| Uktot | 0.149 | Itagen | 0.377 |
| Ukyl | 0.160 | Spatar | 0.383 |
| Ukbir | 0.161 | Itatrie | 0.405 |
| Norway | 0.176 | Fratar | 0.407 |
| Ukmanc | 0.178 | Itavar | 0.417 |
| Scotot | 0.184 | Spamur | 0.421 |
| Scowes | 0.212 | Slovak | 0.424 |
| Itarag | 0.226 | Fradou | 0.428 |
| Finland | 0.231 | Spagra | 0.443 |
| Gerest | 0.256 | Spazar | 0.444 |
| Polwar | 0.261 | Frabr | 0.446 |
| Itafor | 0.311 | Belar | 0.457 |
| Itapar | 0.324 | Frasom | 0.459 |
| Fraise | 0.328 | Fracal | 0.469 |

### DATA ANALYSIS STRATEGY AND TECHNIQUES

Multivariate descriptive statistical procedures were used: Principal Component Analysis[11]; k-means cluster analysis[12 13]; and Kruskal-Wish multidimensional scaling.[14]

The Principal Components are the optimal synthetic descriptors of a multivariate data set.[15 16] The computation of the Principal Components permits the representation of a set of data in terms of new variables (Components), which correspond to the directions of maximal elongation of the data cloud in the space of the original variables. In mathematical terms, the Principal Components are the linear combinations of the original variables allowing the most parsimonious representation (for example, with the minimal number of variables) of the original information. The Components are mutually orthogonal, thus permitting a representation of the original data set devoid of redundancy. The Components are generated in decreasing order of explained variance (Component 1 always explains the highest fraction of variance, and so on). The Components correspond to the independent concepts underlying a given set of data, and their meanings can be rationalised by the inspection of the "factor loadings"—that is, the correlation coefficients of each Component with the original variables. In our case, the 41 male and 44 female tumour types were expressed by their

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.

*Table 3 Factor loadings of tumor profiles. (A) Male loadings, (B) female loadings. The factor loadings are the correlation coefficients between original variables (in this case tumour sites relative percentages) and components, so they help in the elucidation of the meaning of components. In this particular case the components correspond to specific sets of tumours whose relative abundance is correlated at the level of populations*

| *A* | *PCM1* | *PCM2* | *PCM3* | *PCM4* | *PCM5* |
|---|---|---|---|---|---|
| Lip | −0.050 | **0.599** | 0.007 | 0.153 | 0.658 |
| Tongue | **−0.906** | 0.113 | 0.123 | 0.046 | 0.131 |
| Saliv gland | 0.254 | −0.075 | 0.185 | 0.205 | 0.192 |
| Mouth | **−0.888** | 0.034 | 0.141 | 0.036 | 0.062 |
| Oropharynx | **−0.902** | −0.145 | 0.172 | 0.044 | −0.056 |
| Nasopharynx | −0.224 | **0.740** | −0.269 | 0.388 | 0.031 |
| Hypopharynx | **−0.917** | −0.201 | 0.102 | 0.078 | 0.039 |
| Pharynx unsp | **−0.692** | 0.137 | 0.042 | −0.092 | 0.217 |
| Oesophagus | **−0.660** | −0.316 | 0.247 | −0.362 | 0.051 |
| Stomach | 0.280 | 0.576 | 0.133 | 0.094 | −0.117 |
| Small intest | 0.042 | −0.538 | −0.037 | 0.463 | 0.343 |
| Colon | 0.164 | **−0.590** | −0.475 | −0.175 | −0.281 |
| Rectum | −0.002 | −0.253 | 0.209 | −0.384 | −0.146 |
| Liver | −0.290 | 0.476 | −0.363 | 0.454 | −0.232 |
| Gall bladder | 0.265 | 0.539 | −0.154 | 0.392 | −0.277 |
| Pancreas | **0.748** | 0.091 | 0.423 | 0.020 | −0.252 |
| Nose, sinuses | −0.540 | −0.270 | 0.181 | 0.517 | 0.001 |
| Larynx | −0.439 | **0.763** | −0.134 | 0.088 | 0.149 |
| Lung | 0.233 | 0.437 | 0.264 | −0.558 | −0.421 |
| Other thorac | −0.035 | 0.353 | 0.280 | 0.090 | −0.420 |
| Bone | 0.154 | 0.505 | **0.585** | 0.096 | 0.264 |
| Conn tissue | 0.408 | 0.023 | 0.453 | 0.034 | 0.369 |
| Melanoma | 0.482 | −0.565 | 0.113 | 0.160 | 0.141 |
| Breast | −0.159 | −0.191 | −0.213 | 0.367 | −0.332 |
| Prostate | −0.006 | **−0.790** | 0.013 | 0.344 | 0.310 |
| Testis | 0.131 | −0.572 | 0.422 | 0.052 | −0.203 |
| Penis | 0.079 | 0.254 | −0.117 | −0.090 | 0.377 |
| Bladder | 0.034 | 0.194 | **−0.765** | 0.067 | 0.214 |
| Kidney | 0.235 | −0.097 | 0.261 | **0.644** | −0.501 |
| Eye | 0.078 | −0.214 | **0.732** | 0.015 | 0.330 |
| Brain | 0.538 | 0.288 | 0.041 | 0.270 | 0.177 |
| Thyroid | 0.144 | −0.0196 | 0.010 | **0.796** | 0.001 |
| Other endocrine | 0.245 | 0.258 | 0.004 | −0.076 | −0.183 |
| Hodgkin's d | 0.270 | 0.410 | 0.443 | 0.154 | 0.035 |
| non-H lym | 0.262 | **−0.582** | −0.496 | 0.078 | 0.094 |
| Mult myel | **0.640** | −0.326 | −0.477 | −0.009 | 0.211 |
| Lym leuk | 0.337 | 0.258 | 0.210 | −0.327 | 0.240 |
| Myel leuk | 0.516 | −0.026 | −0.263 | −0.291 | 0.287 |
| Oth leuk | −0.204 | 0.008 | −0.200 | −0.024 | −0.067 |
| Leuk unsp | 0.300 | −0.227 | 0.349 | 0.427 | 0.051 |
| % exp var | 18.8 | 16.1 | 9.9 | 8.7 | 6.5 |

| *B* | *PCF1* | *PCF2* | *PCF3* | *PCF4* | *PCF5* |
|---|---|---|---|---|---|
| Lip | **0.660** | 0.079 | 0.048 | 0.332 | 0.251 |
| Tongue | −0.022 | 0.532 | 0.087 | 0.150 | 0.114 |
| Saliv gland | 0.457 | 0.233 | −0.246 | −0.316 | 0.008 |
| Mouth | −0.171 | 0.331 | 0.185 | −0.001 | 0.262 |
| Oropharynx | −0.197 | 0.083 | **0.647** | −0.186 | −0.337 |
| Nasopharynx | 0.389 | 0.332 | 0.023 | 0.043 | −0.057 |
| Hypopharynx | **−0.727** | −0.025 | 0.226 | 0.107 | −0.116 |
| Pharynx unsp | −0.492 | −0.274 | −0.062 | 0.240 | −0.118 |
| Oesophagus | **−0.647** | −0.503 | −0.253 | 0.176 | 0.135 |
| Stomach | **0.564** | −0.061 | −0.181 | 0.251 | −0.097 |
| Small intest | −0.215 | 0.307 | 0.230 | −0.409 | 0.425 |
| Colon | **−0.588** | 0.387 | 0.000 | 0.069 | 0.010 |
| Rectum | 0.257 | 0.070 | **0.598** | 0.246 | −0.094 |
| Liver | **0.684** | 0.357 | −0.427 | −0.162 | −0.235 |
| Gall bladder | **0.732** | 0.189 | −0.159 | −0.408 | −0.162 |
| Pancreas | 0.244 | −0.244 | −0.417 | −0.383 | 0.431 |
| Nose, sinuses | −0.213 | −0.043 | 0.324 | 0.370 | −0.002 |
| Lung | −0.152 | −0.423 | −0.171 | −0.264 | −0.241 |
| Other thorac | −0.476 | **−0.718** | −0.382 | −0.120 | 0.121 |
| Bone | 0.501 | −0.142 | −0.093 | −0.421 | −0.067 |
| Conn tissue | **0.763** | −0.151 | 0.330 | 0.253 | 0.006 |
| Melanoma | 0.533 | 0.092 | 0.476 | 0.305 | 0.007 |
| Breast | −0.380 | 0.193 | 0.443 | −0.061 | **0.598** |
| Uterus unsp | −0.567 | 0.547 | 0.287 | 0.164 | −0.264 |
| Cervix uteri | 0.011 | 0.166 | −0.479 | 0.022 | −0.276 |
| Placenta | 0.517 | **−0.589** | 0.324 | −0.234 | −0.133 |
| Corpus uteri | 0.487 | −0.241 | 0.254 | −0.134 | 0.115 |
| Ovary | **0.788** | 0.247 | 0.326 | −0.084 | −0.039 |
| Oth f gen | 0.404 | **−0.644** | 0.308 | −0.028 | 0.239 |
| Bladder | 0.424 | 0.166 | −0.184 | 0.023 | 0.197 |
| Kidney | −0.503 | −0.012 | −0.415 | 0.212 | 0.283 |
| Eye | 0.121 | 0.129 | 0.327 | **−0.780** | 0.158 |
| Brain | 0.414 | −0.503 | 0.284 | 0.463 | 0.240 |
| Thyroid | 0.709 | 0.203 | −0.302 | −0.170 | 0.059 |
| Oth endoc | 0.120 | **0.670** | 0.188 | −0.220 | −0.036 |
| Hodgkin's d | 0.489 | −0.199 | −0.034 | 0.346 | 0.430 |
| non-H lym | **0.663** | 0.074 | −0.130 | 0.227 | −0.210 |
| Mult myel | −0.472 | **0.699** | 0.003 | 0.059 | 0.122 |
| Lym leuk | −0.219 | 0.526 | −0.460 | 0.077 | 0.406 |
| Myel leuk | 0.517 | 0.123 | −0.010 | 0.406 | 0.066 |
| Oth leuk | 0.278 | 0.197 | −0.395 | 0.302 | 0.139 |
| Leuk unsp | 0.126 | 0.429 | 0.076 | 0.226 | −0.041 |
| % exp var | 22.3 | 11.9 | 9.4 | 7.4 | 5.6 |

first five principal components (PCM1-PCM5 and PCF1-PCF5 for men and women respectively).

The k-means cluster analysis is aimed at highlighting a mathematically optimal partition of the statistical objects (here, regional areas). The optimality criterion of k-means algorithm constructs classes that are the most internally homogeneous (minimal intra-cluster variance) with maximum separation between them (maximal inter-cluster variance).[12 13]

The Kruskal-Wish non-metric multidimensional scaling technique[14] generates two explicit coordinates (axes) in which the points (statistical units) are projected, with the goal of maximising the rank correlation between the original distances among the points and the distances in the projection space. In other words, the algorithm gives an optimal bidimensional representation of complex data sets by maximising the topological resemblance between the "high dimensional" input and the "low dimensional" output.

GLOSSARY OF VARIABLE NAMES

PM: global incidence of tumours per 10000 persons (male)

PF: global incidence of tumours per 10000 persons (female) DELTAN: Gender differential normalised incidence: (PM-PF)/PM PCM1-PCM5: First five principal components of male tumour profile PCF1-PCF5: First five principal components of female tumour profile DEM1-DEM4: First four demographic components

## Results and Discussion

SUMMARISING THE TUMOUR PROFILES OF THE EUROPEAN AREAS INTO NEW DESCRIPTORS

The European areas were described by the profile of induction of 85 different types of tumours (standardised proportion of incidence). These variables were summarised into five Principal Components for the female population (PCF1 to PCF5), and five Principal Components for the male population (PCM1 to PCM5). Principal component analysis is the optimal way of summarising very sparse information (that is, many variables) into a reduced number of new descriptors (Principal Components) that, still conveying the original information, have a more easily manageable size. In the interpretation of the Components, it should be remembered that each of them combines together the information from a set of correlated variables (here, tumour types). Tables 4 (A), (B) report the factor loadings profile of the male and female components respectively. The five Components solution explains 60% and 57% of total profile variability for male and female respectively. These Principal Components were previously demonstrated[10] to be substantially time invariant (average correlations between 1985–1988 and 1988–1992 periods: 0.8–0.9).

Hodgkin's dis = Hodgkin's disease; non-H lym = non-Hodgkin's lymphoma; Mult myel = multiple myeloma; Lym leuk = lymphoid leukaemia; Myel leuk = myeloid leukaemia; Oth leuk = other leukaemias; Leuk unsp = leukaemia unspecified.

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.

*Table 4 (A) Cluster profiles in terms of Principal Components of female tumour spectra. (B) Cluster composition. The prefix of each area code refers to the correspondent country (see table 1)*

| *A* | *Cluster profile* | | | | |
|---|---|---|---|---|---|
| | PCF1 | PCF2 | PCF3 | PCF4 | PCF5 |
| Cluster 1 | 2.98 | −2.42 | 1.11 | 3.78 | 1.21 |
| Cluster 2 | −0.99 | −0.86 | −0.48 | 0.37 | 0.15 |
| Cluster 3 | −0.41 | 0.56 | 1.57 | 0.22 | −0.71 |
| Cluster 4 | −0.05 | 0.55 | 0.16 | −0.95 | 2.11 |
| Cluster 5 | 1.36 | −1.12 | 0.47 | −1.47 | −0.40 |
| Cluster 6 | 0.38 | 0.83 | −0.68 | 0.16 | −0.47 |

| *B* | *Cluster membership* |
|---|---|
| Cluster 1 | BELAR |
| Cluster 2 | DENM, EIRE, SCOTOT, SCOWES, UKBIR, UKMANC, UKMER, UKOX, UKSTH, UKSWR, UKTOT, UKYL |
| Cluster 3 | FRABR, FRACAL, FRADOU, FRAISE, FRASOM, FRATAR, SLOVEN |
| Cluster 4 | FINL, ICELAND, ITATRIE, NORW, SWED |
| Cluster 5 | BOEM, GEREST, POLSLE, POLWAR, SLOVAK |
| Cluster 6 | SPAGRA, GERSAR, ITAFIR, ITAFOR, ITAGEN, ITALAT, ITAPAR, ITARAG, ITATOR, ITAVAR, SPAMUR, SPANAV, SPATAR, SPAZAR |

The interpretation of the Components by the factor loadings (table 3) must be guided by the consideration that the correlations between tumours must be intended solely on population basis. Each person is scored of only one tumour type, thus the correlation between the tumour sites A and B arises from the fact that regional areas having a comparatively high (low) proportion of A have an high (low) proportion of B.

In terms of correlations among tumour types, PCM1 is characterised by the opposition between oral cavity tumours and pancreas plus multiple myeloma tumours. PCM2 is driven by the opposition of lip, nasopharinx, larynx versus colon, prostate, non-Hodgkin's lymphoma. PCF1 is driven by the opposition liver, gall bladder, connective tissue, lip, ovary, non-Hodgkin's lymphoma versus oesophagus, colon, hypopharynx. PCF2 has to do with the balance between genital tumours other than uterus and ovary, non-pulmonary thoracic tumours and endocrine and multiple myeloma tumours. As pointed out above, the associations among tumours are typical of the geographical areas (not of the person); thus any explanation has to be essentially modelled by socioeconomical cultural factors, and only secondarily by biomedical considerations. We plan to perform a refined socioeconomic characterisation of the areas in a future investigation.

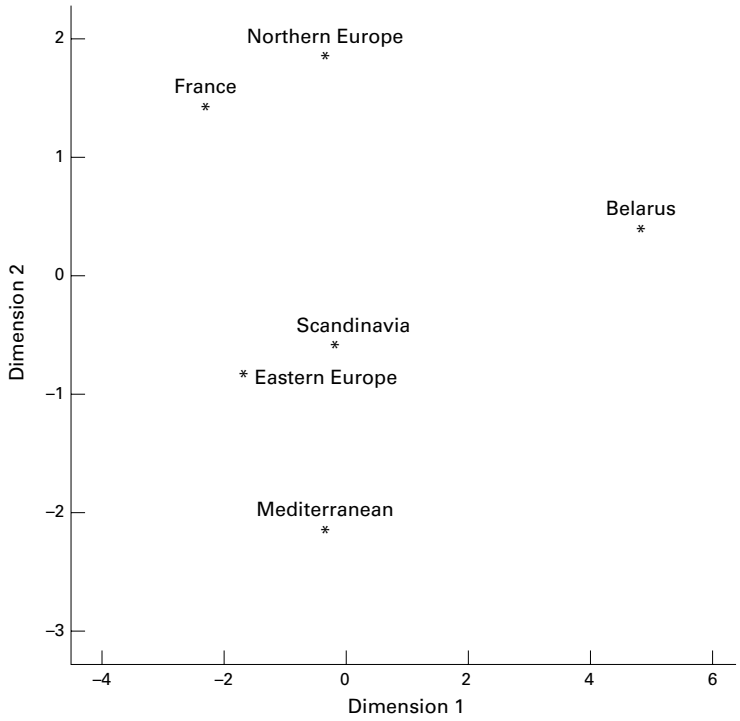


*Figure 1 The space spanned by the first two variables (Dimension 1, Dimension 2) derived by the application of Kruskal-Wish multidimensional scaling procedure to the Euclidean distances among clusters. The location of the points (clusters) in this space is the best (in a least square sense) reproduction of the observed differences in cluster profiles (see table 4B).*

### CLASSIFICATION OF THE EUROPEAN AREAS ACCORDING TO TUMOUR SPECTRA

While planning a detailed interpretation of the Components in terms of socioeconomic and/or environmental determinants, we used the Components for studying the similarity of the European regions in terms of tumour spectra: this was accomplished by applying the k-means cluster analysis to the five female and five male Principal Components separately. The analysis of the female Principal Components pointed to an optimal six classes partition of the 44 areas. This partition is summarised in table 4 and accounts for around 75% of total variability. The clustering of the correspondent male data (PCM1–5) saved around 74% of total variability, and had a relative concordance of 0.88 contingency coefficient (qualitative-case alternative to the Pearson $r$[17]) with the female partition ($\chi^2 = 153.18$; $p<0.001$).

Table 4 shows the concordance between the partition obtained and the classic European "cultural-geographical" areas:

*Mediterranean*: as expressed by Cluster 6 that collect all and only the Italian and Spanish locations (with the only exception of Saarland).

*Eastern Europe*: as expressed by Cluster 5 that collects all and only Eastern Europe areas with the only exception of Belarus that forms an outlier (Cluster 1).

*Scandinavia*: as expressed by Cluster 4 (with the only exception of Trieste that in any case, given its peculiar history of mixed population and habits can hardly be considered homogeneous to the other Italian locations, and is part of Italy only since 1918).

*France*: all the French locations were grouped in the almost "pure French" Cluster 3 (with the only exception of Slovenia that, in any case, is strongly heterogeneous for history and cultural habits to Eastern Europe).

*Northern Europe*: Cluster 2 collected all the United Kingdom locations together with Denmark and Eire.

This result points to the presence of a clear "nation" effect, with the regional areas pertaining to the same nation that cluster together, irrespective of their relative industrial, farming, urban, or rural character and of their relative global incidence of tumours.

It should be noted that this partition has nothing to do with the relative tumour incidence (for example Varese and Ragusa are in the same cluster even if they have a huge (around 50%) difference in cancer average incidence) but only with the relative abundance

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.

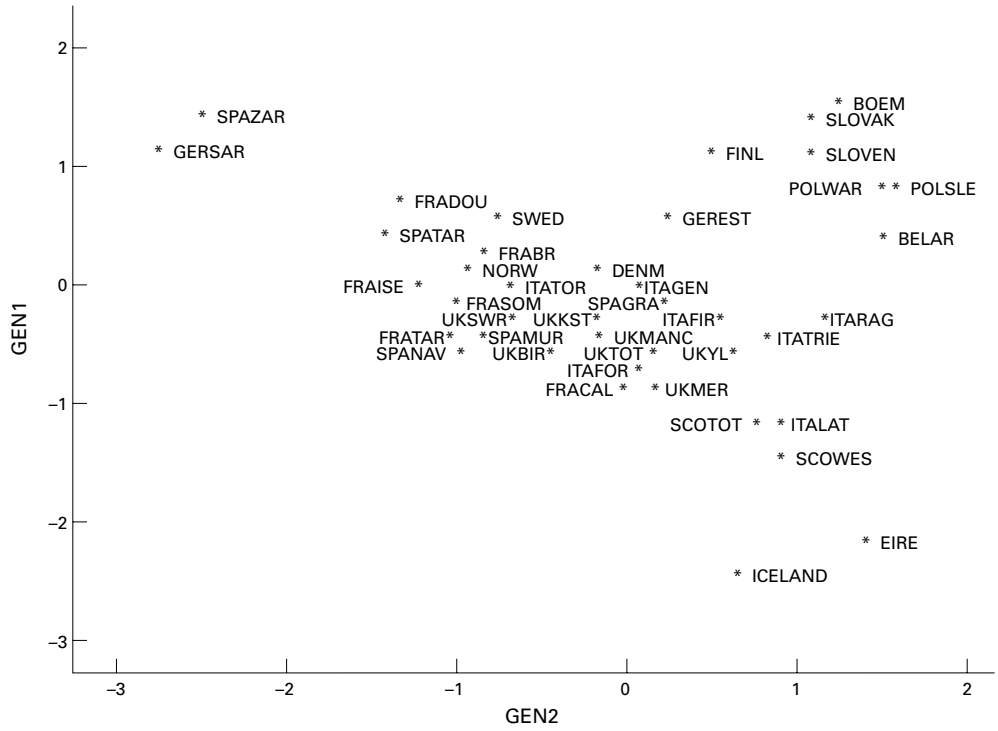


*Figure 2 The space spanned by the two principal components of the blood group (AB*0 system) relative frequencies in the studied areas. The AB*0 system is used as raw estimate of genetic resemblance between populations, based on the fact that it is the only genetic marker for which direct and reliable data are available at the regional level. It is worth noting the intermingled character of European populations that are difficult to separate on genetical basis, with the only possible exceptions of Eastern Europe and small islands (Eire and Iceland).*

of different target sites (this because we normalised for the differences in absolute incidence).

Further results were obtained by applying multidimensional scaling to the between clusters distance matrix: this analysis projected the clusters of areas into a two dimensional plane according to their tumour spectra dissimilarities (distances). Given the high correlation between the male and female partitions, only the female data set was used. Figure 1 shows the presence of "super aggregations" consisting of *France* and *Northern Europe* on the one hand and *Scandinavia* and *Eastern Europe* on the other, while *Mediterranean* and *Belarus* remained distinct poles.

Many hypotheses can be made about these super aggregations: Scandinavia and Eastern Europe are genetically more similar to each other, than to other European areas[3]; moreover, they have been in close contact through migrations and commerce for a long time. Similar arguments can be made for Northern Europe and France. Moreover, it should be remembered that the high genetical flux among Europeans makes Europe very homogeneous from the genetic point of view, and that the genetical distances do not delineate such sharp areas like the one evidenced in this study.[6] As a demonstration of this, we show in figure 2 a map obtained by Principal Component Analysis of AB*0 alleles (blood groups).[18] This genetical map points to a picture much more indistinct than that highlighted by grouping the areas on the basis of tumour spectra. This is a clear indication that "culture" in the broad sense of life style habits has a predominant role in shaping the cancer incidence profiles.

*Table 5 Factor loadings of socioeconomic description. The original variables relevant for the interpretation of components (higher loadings) are in bold. DEM1 explains the major part of variability and is easily interpreted as a "female emancipation" index linked to the economic development of European nations*

| | *DEM1* | *DEM2* | *DEM3* | *DEM4* |
|---|---|---|---|---|
| Population density | 0.046 | 0.046 | **0.912** | −0.051 |
| Migration rate | 0.211 | −0.456 | 0.060 | 0.546 |
| Birth rate | 0.229 | **0.907** | 0.141 | −0.109 |
| Fertlity rate | −0.006 | **0.815** | −0.021 | −0.333 |
| Mother mean age at birth | **0.858** | 0.304 | 0.188 | 0.127 |
| Male mean age | **0.796** | 0.391 | 0.257 | 0.217 |
| Female mean age | **0.888** | 0.182 | 0.194 | 0.140 |
| Infant mortality rate | **−0.829** | 0.353 | −0.093 | −0.115 |
| Urbanisation rate | 0.491 | −0.444 | 0.447 | −0.168 |
| Female occupation rate | −0.305 | −0.105 | **−0.675** | −0.460 |
| Male ocupation rate | −0.073 | 0.200 | −0.005 | **−0.838** |
| GNP per capita | **0.941** | 0.037 | 0.038 | 0.037 |
| Percentage of service industry workers (male) | **0.873** | −0.252 | 0.260 | −0.056 |
| Percentage of service industry workers (female) | **0.864** | −0.074 | 0.215 | 0.209 |
| Percentage of female ministers | **0.821** | 0.050 | −0.280 | 0.032 |
| % expl var | 46 | 20 | 9.9 | 4.9 |

## MALE-FEMALE DIFFERENTIAL INDUCTION OF CANCER

The presence of such a clear correlation structure among the different areas stimulated us to further check it: as a probe, we used a number of socioeconomic variables describing the female condition. The goal of this analysis was to investigate if a "cultural" description of the areas (like that provided by the socioeconomic characterisation of the female condition) was able to explain the observed tumour patterns. Based on the previous results pointing to a nation effect, this confirmatory analysis was performed on a nation basis.

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.

*Table 6 Pearson correlations between socioeconomic component (DEM1) and pathology descriptors. Pearson correlation coefficients between DEM1 and pathology variables (in parentheses, p value under the null hypothesis r=0)*

| *Variable* | *r* | *p Value* |
|---|---|---|
| DELTAN | **−0.742** | (0.001) |
| PM | −0.492 | (0.053) |
| PF | **0.591** | (0.016) |
| PCM1 | 0.379 | (0.148) |
| PCM2 | **−0.843** | (0.0001) |
| PCM3 | **−0.554** | (0.026) |
| PCM4 | 0.474 | (0.064) |
| PCM5 | 0.524 | (0.037) |
| PCF1 | **−0.787** | (0.0003) |
| PCF2 | **0.704** | (0.002) |
| PCF3 | −0.023 | (0.926) |
| PCF4 | −0.247 | (0.355) |
| PCF5 | **0.565** | (0.026) |

Each European country was described by 15 parameters reported in table 5. Principal component analysis (with a subsequent VARIMAX rotation[19]) applied to this socioeconomic data set gave rise to a four components solution (DEM1 to DEM4) reported in table 6. The four components globally explained 81% of total variability, with the first component (DEM1) by itself explaining 46% of total variability. The inspection of the variables maximally loaded on DEM1 shows that DEM1 summarised the advancement in female socioeconomic condition occurring in the past decades in wealthiest countries: in fact the reaching of "apical" positions for women (per cent of female ministers, loading=0.821) goes hand in hand with the per capita GNP (loading= 0.941) and the increase of mother mean age at birth ( loading= 0.858). All this is a consequence of the past decades European history and the inter-nations variability along this component is a comprehensive index of the degree of socioeconomic development.

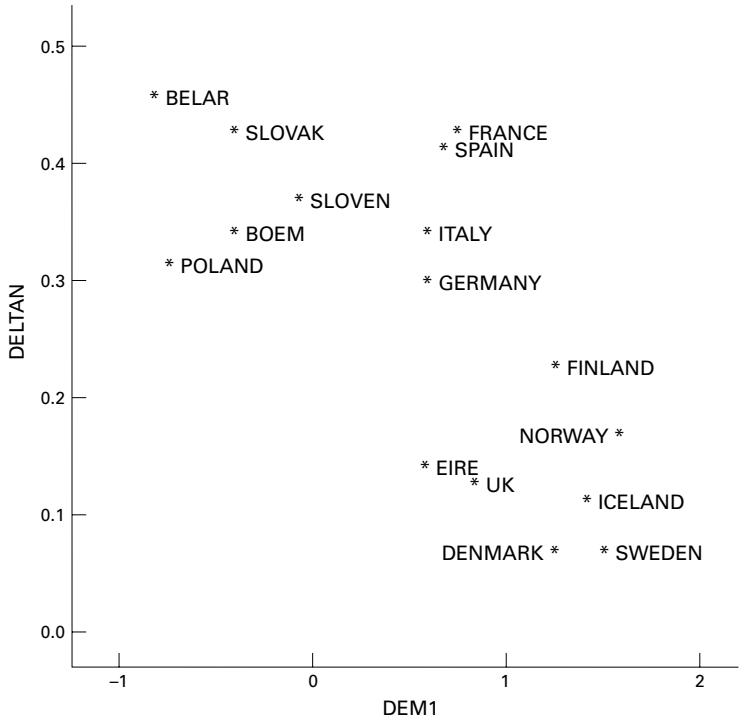


*Figure 3 The relation between DELTAN and DEM1. The negative relation between DELTAN and DEM1 indicates that, in the countries where women emancipation begun earlier (high values of DEM1), the relative differences in tumour incidence between sexes is lower.*

KEY POINTS
- The human environment is shaped by socioeconomic history.
- Environmental factors play a predominant part in cancer induction and determine tumour profiles of the human population.
- The exploitation of sociodemographic/pathology correlations on spatial bases can help to detect yet hidden cancer determinants.
- Multivariate analysis is an invaluable tool for epidemiological research.

Then we averaged the PCM1–5, PCF1–5 and DELTAN data on a nation basis (data not shown). DELTAN is the normalised difference between male and female global tumour incidence. The utility of this parameter is linked to its "pure environmental" character (the biological differences between sexes are obviously identical in all the studied areas and thence the wide between areas variability evident in table 2 is completely attributable to environmental factors) and to the possibility to have a description of the studied areas by means of socioeconomic parameters related to the female condition. Thus the socioeconomic description of Europe under the perspective of female condition should be most naturally comparable with the DELTAN index (in addition to the female tumour profile components).

The only socioeconomic component displaying a statistically significant correlation with the pathology variables was DEM1. The correlations are reported in table 6: it is worth noting the high statistical significance of the relations scored between DEM1 and DELTAN, PCF1, PCF2, PCM2 and PCM3. The presence of a significant correlation between "male" pathology components (especially PCM2) and DEM1 could be puzzling at a first sight, but is important to emphasise that DEM1 conveys information on the whole (and not only female) society development and that PCM2 and PCM3 are in turn strongly correlated with PCF2 (the second female principal component). On the other hand, the first principal component of male tumour profile (PCM1) that does not score any significant correlation with any female component (so being a "pure male" tumour macro cause) is not correlated with DEM1, so confirming the "female condition" character of the socioeconomic component.

Figure 3 shows the relation between DEM1 and DELTAN, and shows how the earlier and more pronounced advancement of female condition in Northern Europe (higher values of DEM1) goes together with a decrease in the gender differences of tumour incidence (lower values of DELTAN).

The quantitative relevance of the obtained results was confirmed by an analysis of variance computed for the nations with a sufficient number of studied regional areas

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.

(France, Italy, Spain, UK). This analysis was aimed at checking the existence of a remarkable "nation effect" for the pathology components correlated with DEM1. This nation effect was effectively demonstrated (DELTAN: $F$=59.5, p<0.0001; PCF1: $F$=21.53, p<0.0001; PCF2: $F$=51.14, p<0.0001; PCM2: $F$=40.78, p<0.0001), so giving both quantitative strength to the observed demography/pathology correlations, and a compelling evidence to the Europe tumour profiles clusterisation.

## Conclusions

This study contributes towards an ecological multivariate approach in epidemiological studies. Its important message is the predominant role of environment—in the complex sense used by ecologists and not in the narrow sense of single toxicologically relevant expositions—in the causation of human cancer. The demonstration of the possibility of modelling the pathology data by socioeconomic descriptors opens the way to further, more refined, analyses aimed at "giving a name" to the tumour profile components, thus enabling the decision makers to undertake practical actions to reduce risks. It should be noted that the relevance of socioeconomic descriptions for understanding pathology profiles is well known to both epidemiologists and pathologists (for example, the epidemiological transition theory as described by Omran[20]).

From a methodological point of view, these results highlight the need for a strong interdisciplinary attitude in dealing with public health problems.

The continued interest of Ann M Richard and Joseph P Zbilut is gratefully acknowledged.

Funding: this publication was supported by the ordinary funding of the Istituto Superiore di Santa' (Government Health Agency).

Conflicts of interest: none.

1 Tomatis L, Huff J, Hertz-Picciotto I, *et al*. Avoided and avoidable risks of cancer. *Carcinogenesis* 1997;**18**:97–105.
2 Taubes G. Epidemiology faces its limits. *Science* 1995;**269**: 164–9.
3 Cavalli-Sforza LL, Menozzi P, Piazza A. *The history and geography of human genes*. Princeton, NJ: Princeton University Press, 1994.
4 Braudel F. *The Mediterranean*. London: Harper-Collins, 1992.
5 Duroselle J-B. *Storia dell'Europa*. Milan: Fabbri, 1990.
6 Piazza A, Rendine S, Minch E, *et al.* Genetics and the origin of European languages. *Proc Natl Acad Sci USA* 1995;**92**: 5836–40.
7 Sokal RR, Oden NL, Rosenberg MS, *et al.* Ethnohistory, genetics, and cancer mortality in Europeans. *Proc Natl Acad Sci USA* 1997;**94**:12728–31.
8 International Agency for Research on Cancer. *Cancer incidence in five continents.* Vol VI. Lyon: IARC, 1992.
9 International Agency for Research on Cancer. *Cancer incidence in five continents*. Vol VII. Lyon: IARC, 1997.
10 Benigni R, Giuliani A. Tumor profiles and incidence in Europe: robustness of spatial patterns of correlation, and their relation with allele frequencies of the AB0 blood group system. *Eur J Epidemiol* (in press).
11 Lebart L, Morineau A, Warwick KM. *Multivariate descriptive statistical analysis*. New York: Wiley, 1984.
12 Everitt B. *Cluster analysis*. New York: Halsted, 1980.
13 Benigni R, Giuliani A. Quantitative modeling and biology: the multivariate approach. *Am J Physiol* 1994;**266**:R1697–704.
14 Kruskal JB, Wish M. *Multidimensional scaling*. Beverly Hills: Sage Publications, 1978.
15 Huang NE, Shen Z, Long SR, *et al*. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc R Soc Lond A* 1998;**454**:743–995.
16 Broomhead DS, King GP. Extracting qualitative dynamics from experimental data. *Physica D* 1986;**20**:217–36.
17 Kendall M, Stuart A. *The advanced theory of statistics*. New York: MacMillan, 1979.
18 Mourant AE, Kopec AC, Domaniewska-Sobzac K. *The distribution of human blood groups and other polymorphisms*. London: Oxford University Press, 1976.
19 Taylor CC. Principal component and factor analysis. In: O'Muircheartaigh CO, Payne CD, eds. *Exploring data structures*. New York: John Wiley, 1979:89–124.
20 Omran AR: A century of epidemiologic transition in the United States. *Prev Med* 1977;**6**:30–51.

J Epidemiol Community Health: first published as 10.1136/jech.54.4.262 on 1 April 2000. Downloaded from http://jech.bmj.com/ on March 7, 2021 by guest. Protected by copyright.