Eur J Epidemiol (2009) 24:381–387
DOI 10.1007/s10654-009-9347-6

INFECTIOUS DISEASES

A graphical study of tuberculosis incidence and trends in the WHO's European region (1980–2006)

Martín Ríos · Toni Monleón-Getino

Received: 6 November 2008 / Accepted: 27 April 2009 / Published online: 19 May 2009
© Springer Science+Business Media B.V. 2009

Abstract A graphical output was obtained using classical principal component analysis techniques in order to analyse tuberculosis trends in Europe over a 27-year period (1980–2006). Taxonomic methods were used to better define the interrelationship between the data in the 52 countries studied. Data were provided by the World Health Organization. Differences in the overall incidence and trends were identified during the 1980–2006 period. The highest rates of incidence were reported in Kazakhstan, Bosnia and Herzegovina, Romania and Kyrgyzstan. High and moderately high rates were reported in the former Soviet Union, the former Yugoslavia, some countries from the former Eastern Bloc, Turkey and Portugal. The lowest rates were reported in the eastern Mediterranean, Scandinavia and Iceland. Risk of infection was determined by social conditions, intravenous drug use, HIV infection and immigration from countries where tuberculosis is endemic. As regards development of tuberculosis in Europe, 1992 represents the change in the decreasing trend in the incidence observed from 1980, when the incidence presented a minimum general trend and started to increase. The linear model calculated to project the rate of increase from 2006 to 2015, reveals the tuberculosis rates observed during the 1980s.

M. Ríos · T. Monleón-Getino (✉)
Department of Statistics, University of Barcelona, Avenida Diagonal 645, 08028 Barcelona, Spain
e-mail: amonleong@ub.edu

M. Ríos
e-mail: mrios@ub.edu

Keywords Tuberculosis · Epidemiology · Europe · Principal component analysis · Numerical taxonomy

Abbreviations

WHO	World health organization
PCA	Principal component analysis
NJCA	Neighbour-joining cluster analysis
UPGMA	Unweighted pair-group method with arithmetic mean

Introduction

Tuberculosis remains a major killer worldwide, with over two million deaths every year. Incidence is exacerbated by socio-economic factors, drug resistance, HIV infection, inconsistent or partial treatment practices and immigration from countries in which the disease is endemic [1–6]. Descriptive epidemiology is used to determine the frequency distribution and extent of the disease. Many studies have focused on tuberculosis epidemiology around the world [1] and in Europe [2–8], indicating that the rate of incidence has increased or decreased to various degrees in different countries. However, this complex research, which often presents a surfeit of information, is unsuitable for use in an objective study focusing on country-to-country similarities and differences and on the evolution of the disease within a given country, especially over a longer time horizon.

At present, large quantities of data must be managed in all fields of medicine. As a result, demand for data processing methods is increasing. Moreover, descriptive statistics and graphical representations of data now make up a significant part of biometry and epidemiology course content [9].

In this study, we analyze the incidence of tuberculosis in Europe by means of a classification tree using taxonomic methods [10] and graphical representations. Data on the number of tuberculosis cases reported over the 1980–2006 period were summarized in order to facilitate the study and track the evolution of the disease in 52 countries.

One of the main objectives of graphical representation is to summarize a large number of variables without sacrificing too much data. This technique is widely used in data analysis and applied problems.

We used principal component analysis (PCA) [11, 12] to display the populations. Small distances in the graph signify similar populations, whereas, large distances signify very dissimilar populations. Representing different populations allows us to carefully interpret the differences between them. In this study, the populations represent 52 countries, with 27 variables considered for each. The variables correspond to the annual incidences of tuberculosis. First, we classified the countries using taxonomic methods. Then, we plotted the output data for these countries as points on a graph in a low-dimensional Euclidean space. In this case, classical PCA was used to obtain the final graphical output.

Methods

Reported cases

The data presented in this study were provided by the World Health Organization (WHO) [13]. As defined in this study, a case of tuberculosis must be officially reported in the country in question. Tuberculosis notification data must be interpreted with caution because of the duration of infection, which affects both incidence and prevalence. However, although the number of reported cases does not always reflect actual incidence, it provides a reasonable estimate of incidence patterns in many developed countries and in some developing countries with effective epidemiological control practices, such as the European countries included in this study. Moreover, the WHO and the European Region of the International Union Against Tuberculosis and Lung Diseases have made recommendations aimed at standardizing the reporting of tuberculosis cases in Europe. The cases in this study are reported in the WHO statistical information system (WHOSIS). WHOSIS is a world reference database bringing together core health statistics from the 193 WHO Member States. It comprises more than 70 indicators, which can be accessed by means of a quick search, by major categories, or through user-defined tables. The data can be further filtered, tabulated, charted and downloaded.

Within WHOSIS, the WHO specifically offers the Global Health Atlas database for tuberculosis, available at: http://www.who.int/topics/tuberculosis/en/. The Atlas is a complete statistics and information service about tuberculosis around the world. For our study we used this service, accessed at: http://www.who.int/globalatlas/dataQuery/default.asp, to obtain the number of reported cases per 100,000 populations in the WHO European region. The database offers information about access and obtaining data, as well as viewing tables, which can be exported to MS Excel files.

The data analyzed correspond to 52 European or bordering countries and one "average" country, each of which has 23 variables corresponding to the declared cases of tuberculosis during the 27-year period under study (1980–2006) [13. 14]. Figure 1 is a map showing the number of reported cases in the European area included in this study, for 2006, calculated using the WHO data. The n-dimensional random vector is $X_h = (X_{h1},\ldots, X_{hn})$ where X_{hi} represents the number of reported cases per 100,000 population in each period $i = 1,\ldots, n$ in the country h. In this study, $h = 52 + 1 = 53$ and $n = 27$.

The missing values in the original WHO data set were replaced using a linear trend algorithm, which replaces missing values with the linear trend for that point. The existing series is regressed on an index variable scaled 1–n. Missing values are replaced with the predicted values. The number of predicted values by the algorithm were as follows: 5 in 1994, 4 in 1993, 3 in 1981–1988, 1991 and 2005–2006; 2 in 1989, 1992 and 2003–2004; and 1 in 1990 and 1996.

Methodology

Each country was assigned the vector $x = (x^1,\ldots, x^n)$ where x^i is the incidence of tuberculosis per 100,000 population in each of the n years divided by 100,000; this can be considered to be the probable number of people in the population with the disease.

The distance between the populations $x = (x^1,\ldots, x^n)$ and $y = (y^1,\ldots, y^n)$ is given by:

$$d = \sqrt{\sum_{i=1}^{n} (x^i - y^i)^2}.$$

The interdistance matrix D is used to obtain graphical outputs through taxonomic analysis [9]. Using taxonomic analysis, similar objects are grouped into clusters (subsets of a set of objects); two clusters may be separate. Based on the distance between pairs of objects, the distance between the new cluster and all other objects is defined. In this study, we used neighbour-joining cluster analysis (NJCA) since it is the most effective way of reproducing the data

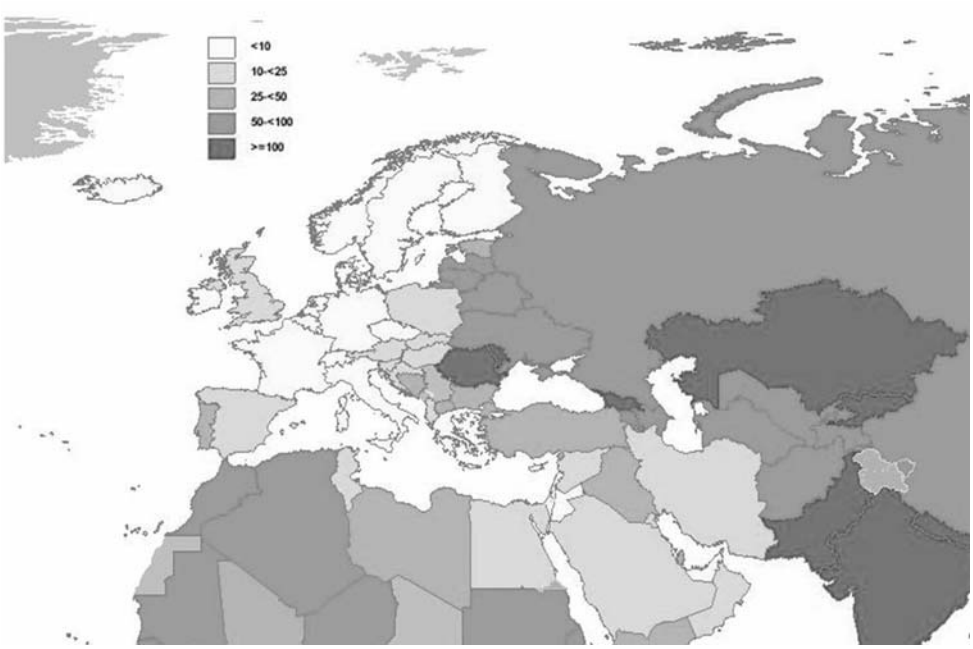

Fig. 1 Map of the 52 countries included in European region covered by this study with the number of reported cases of TB per 100,000 population for 2006, categorized using the WHOSIS database intervals

structure of the two PCA components studied. In addition, NJCA allows study the impact of different countries on tuberculosis incidence, as if it were a phylogeny.

Neighbour-joining cluster analysis [15] is an alternative method for hierarchical cluster analysis. Although it was originally developed for phylogenetic analysis, it may be superior to the unweighted-pair group method with arithmetic mean (UPGMA). In contrast with the UPGMA, in a phylogram (unrooted dendrogram with proportional branch lengths) two branches from the same internal node do not need to have equal lengths and negative branch lengths are forced to zero and transferred to the adjacent branch. The tree is by default rooted on the last branch added during tree construction (this is not midpoint rooting). Optionally, the tree can be rooted on the first row in the data matrix (outgroup).

Other forms of cluster analysis similar to the UPGMA were also used, including paired group, single linkage with Euclidean and Bray Curtis distance and Ward's method [15]. However, the distance between countries as regards tuberculosis evolution and trends could not be reproduced.

Based on the distance obtained, a PCA was performed. This method represents a set of points in n-dimensional space by a subspace of smaller dimensions. The main purpose is to reduce the data from a large number of variables to a smaller number of components, which makes it possible to visualize similarities and differences among the countries studied.

Statistical significance was set at $P < 0.05$. All analyses were performed using SPSS, Version 11 (SPSS Inc., Chicago, Illinois, USA) for PCA and PAST [16] for cluster analysis.

Results

The PCA is summarized in Table 1. All coefficients of the principal components and all eigenvalues are displayed. The inertia percentages accounted for by the first two axes are 77.997 and 13.273%. These components account for more than 91.270% of the total variation. The first component is the "size", with all coefficients positive; it represents the overall incidence during the period studied (total rate of tuberculosis during the period). The first twelve coefficients in the second component are negative, while the rest are positive; this represents the contrast between two periods, 1980–1991 and 1992 onwards. It takes large values for countries with higher incidence of tuberculosis during the second period than during the first. The two-dimensional display obtained is given in Fig. 2.

As shown in Fig. 2, the second PCA component was positive (indicating a higher incidence of tuberculosis during the second period than during the first) in 29 countries (Switzerland, Belarus, Spain, Luxembourg, Turkmenistan, United Kingdom, Iceland, Cyprus, Uzbekistan, Republic of Moldova, Norway, Sweden, Netherlands, Estonia, Bulgaria, Malta, Italy, Israel, Denmark, Ukraine, Lithuania, Russian Federation, Latvia, Armenia, Monaco, Kazakhstan, Romania, Kyrgyzstan and Georgia; incrementally ordered) and negative (higher incidence of

Table 1 Eigenvalues and eigenvectors of the correlation matrix for 27 measurements of the tuberculosis incidence per 100,000 populations in 52 European countries and one average country

Component	First	Second
Eigenvalue	21.059	3.584
Variance that it accounts	77.997	13.273
1980	.852	−.468
1981	.837	−.462
1982	.873	−.444
1983	.884	−.448
1984	.885	−.445
1985	.888	−.436
1986	.902	−.392
1987	.912	−.376
1988	.935	−.319
1989	.930	−.313
1990	.907	−.281
1991	.920	−.221
1992	.806	.032
1993	.789	.190
1994	.872	.077
1995	.925	.118
1996	.934	.228
1997	.857	.327
1998	.896	.361
1999	.905	.372
2000	.900	.397
2001	.905	.397
2002	.887	.431
2003	.871	.454
2004	.873	.424
2005	.843	.427
2006	.836	.436

The eigenvalues are the variances of the principal components; the eigenvectors give the coefficients of the variables

tuberculosis during the firstly period than in the second). It was negative in 24 countries (Bosnia and Herzegovina, Croatia, Portugal, Turkey, Poland, Serbia and Montenegro, Slovenia, Finland, San Marino, Slovakia, Andorra, Greece, Hungary, Czech Republic, Tajikistan, Germany, Albania, Ireland, France, Republic of Macedonia, Belgium, Austria, Azerbaijan and the Average country; incrementally ordered).

Based on the tree, reconstructed by NJCA (Fig. 3), and Fig. 2, there are three large groups with the following subgroups[1]:

Group I-A	Bosnia and Herzegovina, Croatia, Portugal, Poland, Turkey, Serbia and Montenegro, Slovenia, Tajikistan, Hungary, Slovakia and Andorra
Group I-B	Finland, San Marino, Czech Republic, Greece, Germany, Albania, Ireland, France, Belgium and Austria
Group I-C	Spain
Group II	Kazakhstan, Romania, Kyrgyzstan, Georgia, Russian Federation, Lithuania, Latvia, Ukraine, Republic of Moldova, Turkmenistan, Uzbekistan
Group III-B	Azerbaijan, Belarus, Republic of Macedonia, Bulgaria, Estonia, Armenia
Group III-B	Switzerland, UK and Northern Ireland, Luxembourg, Netherlands, Cyprus, Iceland, Sweden, Norway, Malta, Italy, Denmark, Monaco and Israel

The geometric interpretation of the display in Fig. 2 indicates that during the 1980–2006 periods, the countries with the highest reported rates were those in Groups I-A, II and III-A. The countries in Groups I-B and III-B had the lowest values for the first component, which suggests that these countries had the lowest reported rates. The countries with near-average values for the first component were those in Groups I-C and III-A; this suggests that these countries were in the mid-range.

For the second component, the largest positive values correspond to Groups II and III-C, while the lowest positive values correspond to Groups I-A and I-B and moderate high values correspond to Group I-C (Spain). In Group III-A, tuberculosis rates were high, although the second component approached to zero.

Lastly, a graph of the average tuberculosis rate per 100,000 population in the average European country from 1980 to 2006 (Fig. 4) represents the change in the trend beginning in 1992, when the incidence presented a minimum general trend and started to increase. In addition, the linear trend was calculated to project the rate of increase from 2006 to 2015, revealing tuberculosis rates observed during the 1980s.

[1] Ab: Albania, An: Andorra, Ar: Armenia, At: Austria, Az: Azerbaijan, Bl: Belarus, Be: Belgium, Bh: Bosnia and Herzegovina, Bg: Bulgaria, Cr: Croatia, Cy: Cyprus, Cz: Czech Republic, Dk: Denmark, Ee: Estonia, Fi: Finland, Fr: France, Go: Georgia, De: Germany, Gr: Greece, Hu: Hungary, Ic: Iceland, Ie: Ireland, Is: Israel, It: Italy, Ka: Kazakhstan, Ky: Kyrgyzstan, Lv: Latvia, Lt: Lithuania, Lu: Luxembourg, Mt: Malta, Mo: Monaco, Nl: Netherlands, No: Norway, Pl: Poland, Pt: Portugal, Rm: Republic of Moldova, Ro: Romania, Ru: Russian Federation, So: San Marino, Sm: Serbia and Montenegro, Sk: Slovakia, Sl: Slovenia, Es: Spain, Se: Sweden, Ch: Switzerland, Ta: Tajikistan, Yu: The former Yugoslav-Republic of Macedonia, Tr: Turkey, Tu: Turkmenistan, Ur: Ukraine, Uk: United Kingdom, Uz: Uzbekistan, Av: Average country.

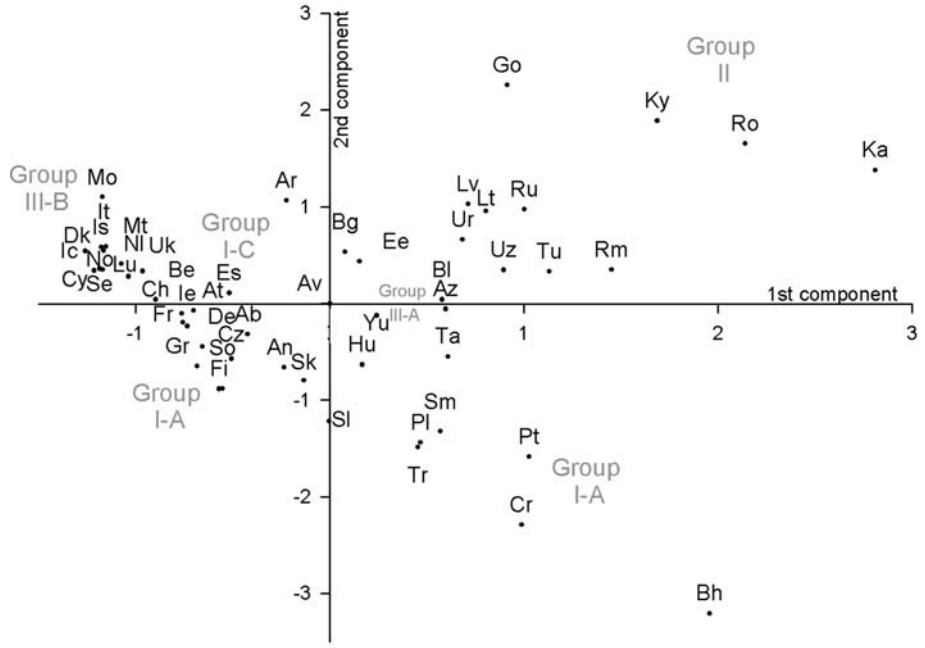

Fig. 2 Two-dimensional graphical display of 52 European countries and one average country, based on the PCA. The two dimensions account for 91.3% of the inertia (Ab: Albania, An: Andorra, Ar: Armenia, At: Austria, Az: Azerbaijan, Bl: Belarus, Be: Belgium, Bh: Bosnia and Herzegovina, Bg: Bulgaria, Cr: Croatia, Cy: Cyprus, Cz: Czech Republic, Dk: Denmark, Ee: Estonia, Fi: Finland, Fr: France, Go: Georgia, De: Germany, Gr: Greece, Hu: Hungary, Ic: Iceland, I.e: Ireland, Is: Israel, It: Italy, Ka: Kazakhstan, Ky: Kyrgyzstan, Lv: Latvia, Lt: Lithuania, Lu: Luxembourg, Mt: Malta, Mo: Monaco, Nl: Netherlands, No: Norway, Pl: Poland, Pt: Portugal, Rm: Republic of Moldova, Ro: Romania, Ru: Russian Federation, So: San Marino, Sm: Serbia and Montenegro, Sk: Slovakia, Sl: Slovenia, Es: Spain, Se: Sweden, Ch: Switzerland, Ta: Tajikistan, Yu: The former Yugoslav-Republic of Macedonia, Tr: Turkey, Tu: Turkmenistan, Ur: Ukraine, Uk: United Kingdom, Uz: Uzbekistan, Av: Average country). The graph includes a classification of the groups based on PCA quadrants: Group I-A: 1st component high, 2nd component low; Group I-B: 1st component low, 2nd component low; Group I-C: 1st component low, 2nd component positive but near zero; Group II: 1st component high, 2nd component high; Group III-A: 1st component high, 2nd component (negative or positive) near zero; Group III-B: 1st component low, 2nd component high

Discussion

Main findings

In this study, graphical representation was used to plot the data for the various countries, based on the rates of tuberculosis incidence. This decision was fortunate because all coefficients of the first component clearly represent the "size", i.e., the overall tuberculosis incidence rates during the period studied. The second axis has increasing coefficients, with the first twelve negative and the last eighteen positive. This axis represents the behavior of the evolution of the trend, i.e., those countries represented by points whose second component has a high positive value are interpreted as indicating that tuberculosis notification rates have risen in recent years. Conversely, if the coordinates of the points have a lower-value second component, this is interpreted as indicating that tuberculosis notification rates have declined in recent years.

This regular pattern—all positive values for the coefficients of the first axis and increasing values for the coefficients of the second principal axis—suggests that tuberculosis incidence has evolved uniformly in the WHO's European region. The taxonomic method was used to group the various countries.

This study shows differences in tuberculosis incidence in various countries and regions of the WHO's European region. Based on the tree and principal component representation, the countries of the newly independent states (NIS) of the former Soviet Union (except Armenia); the former Yugoslav-Republic of Macedonia central and eastern Europe (except Albania and Czech Republic); Turkey; and Portugal, were those in which the largest number of cases was reported during the 1980–2006 period. Over the years, this situation has deteriorated in Romania, the Baltic States, Bulgaria and the NIS (except Azerbaijan, Republic of Moldova, Uzbekistan, Belarus and Ukraine), where the incidence of the disease has remained constant. In recent years, the situation has improved in Tajikistan, central and eastern Europe (except Romania and Bulgaria), the former Yugoslavia, Turkey and Portugal. Certain Mediterranean countries (Malta, Cyprus, Italy and Israel), together with Norway, Sweden, Denmark, Iceland, the Netherlands, Luxembourg and the United Kingdom, had a low incidence,

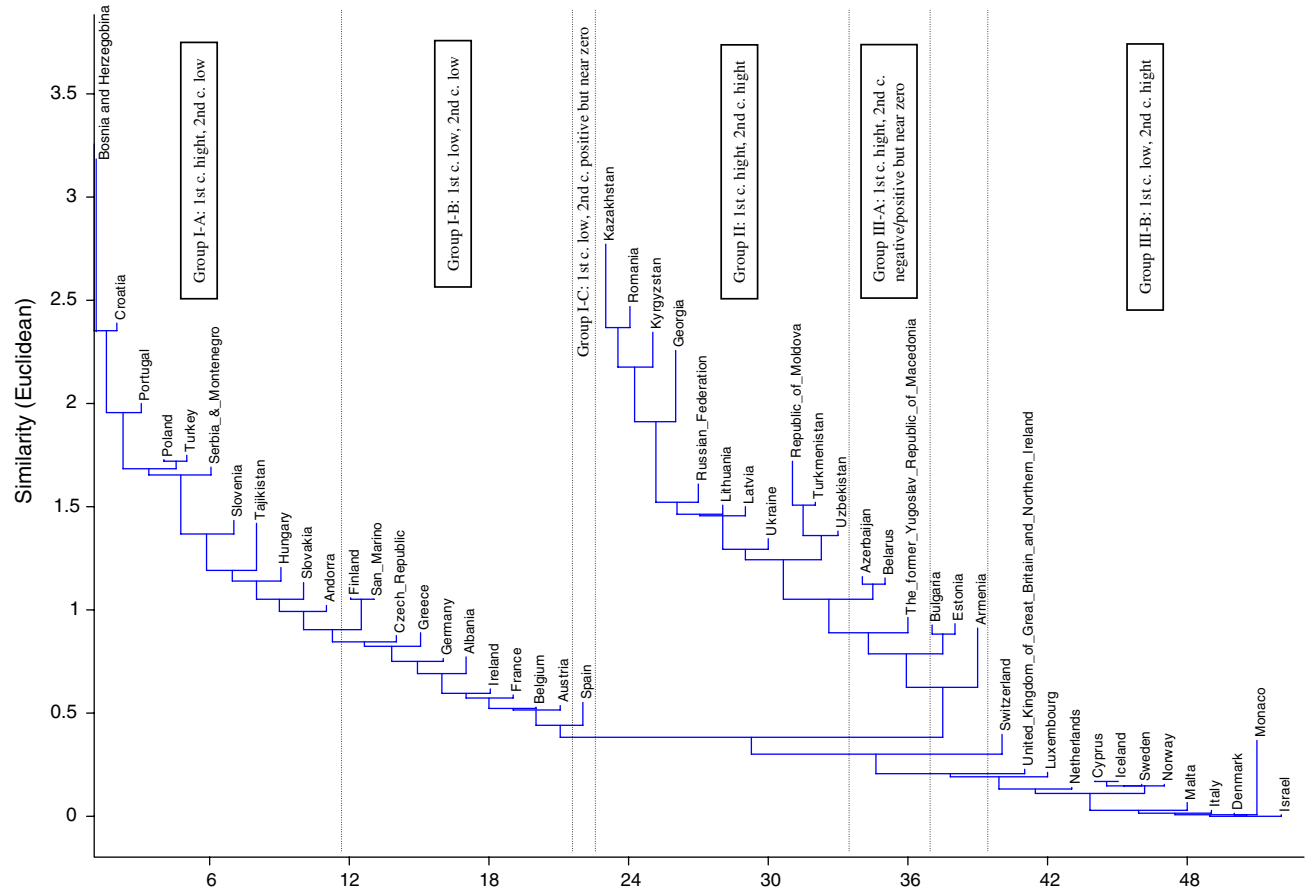

Fig. 3 Cluster analysis using the neighbour-joining cluster analysis method and classification of groups by tuberculosis incidence. The groups assigned to clusters and based on PCA results were, Group I-A: 1st component high, 2nd component low; Group I-B: 1st component low, 2nd component low; Group I-C: 1st component low, 2nd component positive but near zero; Group II: 1st component high, 2nd component high; Group III-A: 1st component high, 2nd component (negative or positive) near zero; Group III-B: 1st component low, 2nd component high

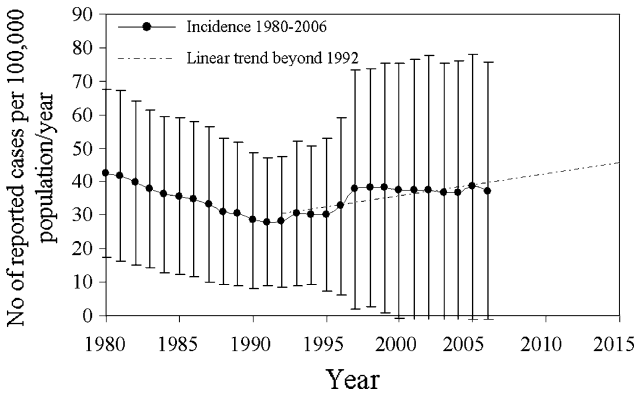

Fig. 4 Graphical representation of average European country tuberculosis incidence and its standard deviation (*line with circular points*). The lineal tendency of the incidence beyond 1992 is plotted (*dotted line*) and extrapolated until 2015

although it has risen in recent years. Austria, Belgium, France, Ireland, Germany and Switzerland had a moderate and constant incidence. Marking a significant improvement, Greece had a moderate incidence. Also showing improvement were Albania, the Czech Republic and Finland, all of which had a moderate incidence. Spain and Armenia had a moderate overall incidence, although it has risen in recent years.

Existing knowledge on this topic

Socio-economic status, together with increased unemployment, homelessness, alcoholism, HIV transmission and drug resistance, boost the incidence of the tuberculosis [14, 17]. This trend is particularly significant in the countries of central and Eastern Europe and the NIS. Political and armed conflicts in the former Yugoslavia also contributed to this trend. Movements of people from countries with a high incidence of tuberculosis, drug abuse and HIV are associated with tuberculosis incidence in Western Europe, especially Spain [18], where the situation has deteriorated during the last years. Portugal has the highest overall

incidence in Western Europe, although significant strides have been made recently. The situation has also improved in the countries of central and eastern Europe (except Romania and Bulgaria, as well as the former Yugoslavia in the post-war period). The Global Health Atlas for tuberculosis offers a complete report on the general epidemiology, strategy, and financing of global TB Control. It contains data from 196 countries and provides information on the trends and scale of TB epidemics; implementation and the impact of the Stop TB Strategy; and progress towards the Millennium Development Goals [13].

Contribution of this study

Every year, reports on the incidence of tuberculosis in Europe are issued. These reports are very thorough and are very helpful to those seeking information on the incidence of the disease in a given country in a given year. However, given the enormous amount of quantitative data compiled on so many countries over the years, it becomes very difficult to track the evolution of the disease and to analyze the similarities and differences between countries in this regard (incidence and evolution).

This study provides a summary of data corresponding to a large number of variables (tuberculosis incidence in Europe over a 27-year period in 52 countries) using a single standard graph, in addition to a dendrogram. The graphical output facilitated the study of tuberculosis's evolution in various countries. Using these graphical methods, we were able to study and interpret tuberculosis incidence in Europe in a straightforward and objective way. This study also aims to encourage future researchers to present information accompanied by methods of summary representation.

References

1. Raviglione MC, Snider DE, Kochi A. Global epidemiology of tuberculosis. Morbidity and mortality of a worldwide epidemic. JAMA. 1995;273:220–6. doi:10.1001/jama.273.3.220.
2. Dye C, Watt CJ, Bleed DM, Mehran S, Raviglione MC. Evolution of tuberculosis control and prospects for reducing tuberculosis incidence, prevalence and deaths globally. JAMA. 2005;293:2767–75. doi:10.1001/jama.293.22.2767.
3. Raviglione MC, Sudre P, Rieder HL, Spinaci S, Kochi A. Secular trends of tuberculosis in Western Europe. Bull World Health Organ. 1993;71:297–306.
4. Rieder HL. Epidemiology of tuberculosis in Europe. Eur Respir J. 1995;8:620s–32s.
5. Sutherland I, Styblo K, Sampalik M, Bleiker MA. Annual risk of tuberculosis infection in 14 countries, derived from the results of tuberculin surveys in 1948–1952. Tuberculosis Surveillance Research Unit Report N.2. Bull Int Union Tuberc. 1971;45:75–114.
6. Waaler II, Gattung O, Mordal K. The risk of tuberculosis infection in Norway. Tuberculosis surveillance research unit report N.2. Bull Int Union Tuberc N.3. 1975;50:5–61.
7. Raviglione MC, Rieder HL, Styblo K, Khomenko AG, Esteves K, Kochi A. Tuberculosis trends in Eastern Europe and the former USSR. Document WHO/ITB/94.176, World Health Organ, 1994.
8. Office fédéral de la santé publique. La tuberculose en Suisse 1988–1992. Bulletin de l'Office fédéral de la santé publique. 1993;41:739–45.
9. Haux R, Dudeck J, Gaus W, et al. Recommendations of the German association for medical informatics, biometry and epidemiology for education and training in medical informatics. Methods Inf Med. 1992;31:60–70.
10. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958;28:1409–38.
11. Anderson TW. An introduction to multivariate statistical analysis. New York: John Wiley & Sons; 1958.
12. Muirhead RJ. Aspects of multivariate statistical theory. New York: John Wiley; 1982.
13. Report WHO. Global tuberculosis control epidemiology, strategy, financing. Geneva. World Health Organization. 2009. WHO/HTM/TB/2009.411 (Available at: http://www.who.int/tb/publications/global_report/2009/en/index.html). Accessed 23 Mar 2009.
14. Report WHO. Global Tuberculosis Control. Interim Policy on Collaborative TB/HIV Activities. Geneva. World Health Organization. 2004. WHO/HTM/TB/2004.
15. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25.
16. Hammer O, Harper DAT, Ryan PD. PAST: palaeontological statistics software package for education and data analysis. Palaeontologica Electronica. 2001;4(1):9.
17. Centers for Disease Control and Prevention. Tuberculosis outbreak among homeless persons, King County Washington, 2002–2003. MMWR Morb Mortal Wkly Rep. 2003;52:1209–10.
18. Mayoral JM, Garcia M, Varela MC, Fernandez JC, Garcia J, Herrera D, et al. Incidente of pulmonary tuberculosis and HIV coinfection in the province of Seville, Spain, 1998. Eur J Epidemiol. 2001;17:737–42. doi:10.1023/A:1015632013465.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.