AGE AND GENDER DEPENDENCE AND THE SECULAR TREND OF DEATH CAUSES IN THE U.S., ANALYSED ON DIVERSITY CURVES

Rudolf Izsák – János Izsák

ABSTRACT

Age vs. death cause diversity curves seem to be the appropriate objects to study death cause diversity. The findings of the present paper are consistent with earlier findings of ad-hoc methodological pilot studies. The paper serves as a significant reference for further studies on epidemiological diversity or concentration. The diversity of death causes has become an important population character to be investigated recently. Its variations treated in the paper may correlate with other demographic characteristics of the population. Furthermore, the revealed variations in death cause diversity refer to the need for standardisation before comparing death cause diversities of various populations.

Keywords: diversity, concentration, death causes, age dependence, sex differences, secular changes

Rudolf Izsák
Department of Algorithms and their Applications, Eötvös Loránd University
E-mail: irudolf@elte.hu

János Izsák
Department of Systematic Zoology and Ecology, Eötvös Loránd University
E-mail: ijanos@staff.elte.hu

INTRODUCTION AND OBJECTIVES

The frequency distribution of death causes can be regarded as an essential component of the demographic features of a population or subpopulation. Specifically, the concentration of death causes is an epidemiological or demographic indicator that correlates with given environmental and health care conditions or age structures. Besides, variations in the mortality and morbidity features of the death cause frequency may refer to future trends in mortality and may have implications for the adequacy of health policies. In economic terms, there is an observable correlation between diagnosis concentration and mean medical expenses concerning various patient groups. At any rate, "high CoD (causes of death) diversity implies major challenges to health-care sysems" (Calazans and Permanyer, 2023).

Numerous investigations have been done concerning the diversity of death causes in several countries, regarding given years. We aim to demonstrate the age and sex specificity of death cause distributions – or to put it differently, death cause diversity and the secular tendency in its change. We use death cause statistics from the United States (USA) regarding the years 2001 and 2006. Some possibilities for the application of the observed diversity conditions will also be presented. Furthermore, we discuss the possibility of applying some other diversity measures and offer new perspectives for further diversity studies on death causes.

LITERATURE

Detailed studies on the diversity or concentration of death causes and morbidity diagnoses are relatively rare in the existing literature. Herdan (1957) was the first to discuss the diversity of diagnoses with regard to hospitalized patients. In the 1970s, new application possibilities for the entropy of diagnoses emerged. Evans and Walker (1972) revealed that in a multivariate linear regression model, the "cost/treated case" ratio parameter (the response variable) correlates with the "complexity parameter" (the explanatory variable). The latter complexity parameter, in turn, bears a relation to the diagnosis entropy (see e.g. Hornbrook, 1982a, 1982b; Farley, 1989; Farley and Hogan, 1990; Lasser et al., 2006). Also citing this result, some authors have recently reported observations on the entropy of morbidity diagnoses referring to a large sample of hospitals (Ostermann and Schuster, 2015).

Recently, Bergeron-Boucher et al. (2020) investigated the diversity (diversification) of mortality causes in entire disease sections in numerous countries. McMonagle et al. (2022) performed similar investigations on the diversity of death causes among subpopulations within Scotland. In both papers, citing our former results, the Shannon index (see below) is applied as a diversity measure.

Another kind of death cause diversity was recently reported on mortality statistics of the United States (Trias-Llimós and Permanyer 2023).

In our own investigations, first we carried out death cause diversity analyses in relation to several disease sections on Hungarian mortality statistics for the years 1967 and 1975, based on ICD-7 and ICD-8, respectively (Izsák and Juhász-Nagy, 1981–1982). Later, we studied the Vital Statistics from the US, Section 1, General Mortality, Deaths, 281 causes (Izsák and Juhász-Nagy, 1984). This study was performed for various, at that time racially defined groups for 1974 and 1975. Numerous diversity indices were applied in this study, among others, the Shannon index, Simpson's reciprocal diversity index, the Gini-Simpson index, both related to the fractionalization index mentioned by Calazans and Permanyer (2023), and some members of Hill's and Hurlbert's index families. In a further study, we found certain main characteristics also relating to age dependence, sex differences and the annual trend of the concentration of mortality cause diagnoses, analyzing circulatory diseases and neoplasms with regards to several countries, such as England and Wales, Norway, Finland, Hungary and Japan (Izsák, 1986, 1993).

The question arises as to why diversity indices are commonly applied instead of concentration indices. As known, diversity and concentration are in a sense opposite terms (Izsák, 1993; Ostermann and Schuster, 2015). That is, the increase in diversity is equivalent to the decrease in concentration. Therefore, most statements on trends in diversity imply the opposite statement regarding concentration.

DATA AND METHODS

Data and their availability

The data sets supporting the conclusions of this article are available in the vital statistics of the United States for the years 2001 and 2006 (NCHS, 2021). They include data on the total number of deaths for each cause by five-year age groups between 1999 and 2006 (GMWKI). Tables include case numbers for most disease categories for 5-year age groups by sex and "race". These statistics have been chosen due to their standard quality, large case numbers and easy availability.

The tenth revision of the International Classification of Diseases (ICD-10) was used (see Moriyama et al., 2011).

Notwithstanding its usefulness, the utilized statistical data has its limitations. Firstly, the statistical tables in question are demographic (static) tables, because the underlying statistics are not based on equilibrium populations or cohorts. This can be crucial when considering the age dependence of some outlined properties. Moreover, the ICD categories are not so clear-cut entities as are e.g. the species in biological taxonomy. Furthermore, the designation of the underlying cause of death is, in some cases, vague. Numerous shortcomings of epidemiological and demographic statistics are widely known. At the same time, the production and the study of such statistics, including mortality statistics, is unavoidable, despite these apparent shortcomings. Similar situations occur in other statistical fields as well.

In similar analyses, the authors generally do not appropriately specify the set of death cause categories selected for the study. This can be attributed to some of the above-mentioned problems in the case of numerous ICD death cause categories. Namely, a considerable part of the death causes are classified as others, ill-defined, or unspecified, etc. In addition, some causes in statistical source data, partly due to extremely small case numbers, are combined. For these reasons, it is reasonable to omit or combine some categories of ICD death causes in certain investigations. In such cases, we found it reasonable to combine a frequent cause with a less frequent one, regardless of being rather similar.

All these conditions make the delimitation of the study set vague to some degree. Even so, only the leading or the largest case numbers have a crucial impact on the values of the most frequently used diversity or concentration indices, such as the Shannon index or the Simpson indices. In addition, these frequent causes are generally only slightly affected by the above-mentioned uncertainties. For example, in the neoplasm disease section, the causes with total case numbers exceeding 1000 were regarded as "leading causes" and were taken into account as such. Our former numerical experimentations affirm this decision. Concerning similar methodological issues, we refer to our earlier findings (Izsák, 1989).

We demonstrate the above considerations with a concrete example taken from the 2001 US mortality statistics (whites, males). In the section of *neoplasm diseases,* we omitted C00 (lip, case number 3) and took it into account in a combined form [C01 (base of tongue 96) + C02 (other parts of mouth 918)]. The further causes of death taken into account separately or combined: [C10 (oropharynx 305) + C11 (nasopharynx 269) + C12 (pyriform sinus 82) + C13 (hypopharynx 123)], separately

C14 (other sites … pharynx 958), C15 (oesophagus 8205), C16 (stomach 5568), C17 (small intestine 462), C18 (colon 20026), [C20 (rectum 3157) + C21 (anus and … 164)], C22 (…), etc. The final cause taken into account is the combination [C94 (other leukaemias 45) + C95 (other leukaemia… 1192)].

In the case of the *circulatory diseases* section, the first causes up to cause I09 (other rheumatic heart disease, with case number 166) were omitted; causes I10 (essential (primary) hypertension 2986), I11 (hypertensive heart disease 7395), I12 (hypertensive renal disease 2365) were separately taken into account; I20 (angina pectoris 140) was omitted; I21 (acute myocardial infarction 85 280) were taken into account, etc. We delimited the set of causes similarly with regard to the *respiratory section* of diseases.

Entropy as a measure of diversity

A number of diversity measures, in the sense of inverse concentration measures, can be used. Taking into account that no essential differences in diversity tendencies were found between applying different diversity indices in former studies (Izsák, 1993, 2005), only the most frequently used diversity index, the Shannon entropy H is applied in the current study. The Shannon entropy has favourable statistical properties. It was the preferred index, in essence e.g., by Bergeron-Bousher et al. (2020) and McMonagle et al. (2022). Moreover, the interpretation of H in *information theory* is straightforward. The formula of H and its applied simple standard estimate, $\hat{H}$ is

$$H = -\sum_{i=1}^{s} p_i \log p_i \approx \hat{H} = -\sum_{i=1}^{s} \frac{n_i}{n} \log \frac{n_i}{n},$$

where p_i is the (positive) occurrence probability of category i; s is the number of categories, n_i is the case number regarding category i (=1,..., s), n is the sum of the n_i's. The base of the logarithm is practically unessential – base 10 logarithms are used in this study. The value of should be considered zero in case n_i=0.

RESULTS AND DISCUSSION

Age dependence of death cause diversity

To demonstrate the results, age–diversity curves were applied. These curves demonstrate the features of age dependence, sex differences and secular trends.

The age dependence of death cause diversity can be established regarding both sexes in all studied disease sections, except for some lower and upper age groups with small case numbers. Namely, a "hollow" starts evolving at about the age intervals 30–34 or 35–39, in most cases, resulting in a diversity minimum (see *Figures 1–4*). This phenomenon emerges in all studied disease sections. The hollow can be attributed to socio-economic factors (Izsák, 1993; Lasser et al., 2006) with a maximum impact seen at age 60–75. It can be linked to the trends in leading causes of death among the elderly in the United States (Gorina et al., 2006).

The relevance of factors such as diet, lifestyle, smoking, etc., is evident. This concept is also supported by the fact that in earlier studies, the hollow was practically missing for African Americans (c.f. with the case studies of "low mortality" population groups by McMonagle et al., 2022). As the diversity curves demonstrate, the intensity of the hollow formation differs with regards to the various disease sections.

Figure 1: Death cause diversity for circulatory diseases by age group and sex, USA, 2001 and 2006

Source: NCHS, 2021.

Figure 2: Death cause diversity for neoplasms by age group and sex, USA, 2006

Source: NCHS, 2021.

Sex differences

In some disease sections, a clear excess of either male or female diversity is observable in the older age groups – for example, in the case of circulatory diseases *(Figure 1)*. In the case of neoplasms, higher female diversity excess is confined to the upper age groups. In mid-life, between the age groups of 30–34 and 55–59, a male diversity excess is apparent *(Figure 2)*. In the case of respiratory diseases, there is not a clear diversity difference between sexes, except for a male diversity excess in the first age groups and a female diversity excess in some upper age groups *(Figure 3)*.

Figure 3: Death cause diversity for respiratory diseases by age group and sex, USA, 2006

Source: NCHS, 2021.

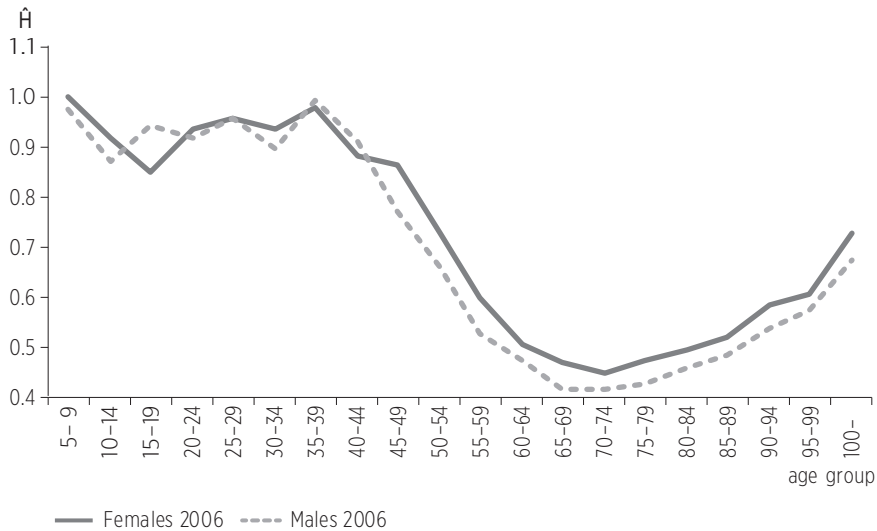
Secular changes in diagnosis entropy

In light of the socio-economic background of the diagnosis entropy or diversity differences, the observed secular trends are not surprising. Furthermore, social campaigns and medical policies usually address the leading death causes. These imply a reduction in the extreme dominance or concentration of the leading death causes.

For example, such a trend can be observed in the case of circulatory diseases in the middle age groups, where diversity increased between 2001 and 2006 regarding both sexes (see *Figure 1*). At the same time, diversity decreased in the case of respiratory diseases among females between 2001 and 2006 *(Figure 4)*.

Furthermore, we can observe e.g. a horizontal dislocation of a given section of a curve expressing a time delay in the corresponding male or female diversity.

Figure 4: Death cause diversity for respiratory diseases by age group, females, USA, 2001 and 2006



Source: NCHS, 2021.

CONCLUSIONS AND PERSPECTIVES

Characteristic variations of entropy by age groups, sex and calendar year of death can be established. They include the characteristic hollow observable in the age-diversity curves in all studied disease sections, the frequently appearing diagnosis diversity excess in the female groups, and the secular increase in diversity with regards to certain disease sections.

The discussed findings are instructive global epidemiological observations on the somewhat underutilised death cause statistics, even bearing the latter's statistical weaknesses. The introduced diversity curves are simple, efficient and convenient objects when investigating the diversity of death causes in various populations.

The proliferation of epidemiologic data sets worldwide, even if carrying numerous uncertainties of different nature, offers an opportunity to search for new epidemiological findings. The initiation of new studies can result in novel demographic and epidemiologic discoveries. One such example can be the present work, reporting characteristic trends of death cause diversity. Namely, the age–diversity graphs in the studied death cause statistics have

a minimum at age 60–75. Furthermore, women tend to show an excess of diagnosis diversity. Finally, in some sections, a unanimous secular change in diagnosis diversity is observable. Most of these findings can be ascribed to certain socioeconomic factors: on some trends in the therapy of major death causes, see Gorina et al. (2006). Additionally, some aspects of the sensitivity of cause-specific death hazards regarding changes in certain factors were recently discussed by Caswell and Ouellette (2016).

Some possible applications of the findings could be the further study of the relationship between the diversity of causes and hospital costs.

Beyond the above points, there are a few additional implications. As mentioned, the entropy index *H* has a straightforward interpretation in information theory. However, when using other diversity indices, further diversity interpretations may also emerge. For example, such measures are the *s(m), m* = 2, 3,... members of Hurlbert's diversity or rarefaction index family (Hurlbert, 1971). In the present case, these indices express the mean number of death causes occurring in a sample consisting of m randomly and independently selected concrete death causes. The formula of the index family and the unbiased estimates using the above symbols are:

$$s(m) = \sum_{i=1}^{s} (1 - (1 - p_i)^m),$$

$$\hat{s}(m) = \sum_{i=1}^{s} \left(1 - \binom{n - n_i}{m} \Big/ \binom{n}{m} \right) \qquad m = 2, 3, \ldots$$

(It should be noted that the sensitivity properties of index (10) are similar to those of the Shannon index (Izsák, 2007), as shown by the corresponding curve in *Figure 1.*)

The index *s*(2) is related to the average number of switches from the "type" of a concrete diagnosis to the "type" of the next one in a series of cases. Such a switch can be linked to a diagnose-specific cost of switching to new medical preparations. Then, the *s*(2) diagnosis diversity or the 1/*s*(2) diagnosis concentration can be linked to the cost component of the medical procedures. More specifically, suppose we know the cost of a single *switch* or change in a series of medical procedures (in the sequence *a a a b b c c a* there are three switches, namely, between (the third) *a* to *b*, *b* to *c* and from *c* to *a*). The introduction of an index expressing the average number of switches in a random series of

procedures as a function of diversity/entropy will make it possible to reveal a concrete link between diversity and a certain cost component. An antecedent of such analyses is published by Brindle and Gibson (2008), discussing the expenditures of the application of some medical devices. As the authors formulate, "... a reduction in diversity is associated with a lower risk of incidents, simplified maintenance and a lower training burden ..." (Brindle and Gibson, 2008: 399).

Another prospective study could be to carry out similar studies on morbidity diagnoses originating from data on, e.g., hospitalized inpatients or results of general screenings. We formerly investigated similar age vs. *H* diversity curves in the case of Hungarian all-diagnosis morbidity, observing approximately maximum curves.

Finally, it should be mentioned that after postulating taxonomic or similar differences or distances of diagnosis categories, one could apply the so-called *quadratic diversity indices* used in mortality and morbidity diversity studies, given that these indices can express the average difference between two randomly chosen diagnoses. Such analyses may significantly widen the realm of diversity studies on epidemiologic statistics.

ACKNOWLEDGEMENTS

We are grateful to the reviewers for providing numerous valuable pieces of advice concerning, among others, methodological issues. Numerous useful corrections were suggested by the Editors.

REFERENCES

Bergeron-Boucher, M.-P., Aburto, J. M. and van Raalte, A. (2020). Diversification in causes of death in low-mortality countries: Emerging patterns and implications. *BMJ Global Health,* 5(7), e002414.

Brindle, G. W. and Gibson, C. J. (2008). Entropy as a measure of diversity in an inventory of medical devices. *Medical Engineering and Physics,* 30(3), 399–401.

Calazans, J. A. and Permanyer, I. (2023) Levels, trends, and determinants of cause-of-death diversity in a global perspective: 1990–2019. *BMC Public Health*, 23(1), 1–12.

Caswell, H. and Ouellette, N. (2016). Mortality and causes of death: matrix formulation and sensitivity analysis. Paper presented at the European Population Conference. http://epc2016.princeton.edu/uploads/160437

Evans, R. G. and Walker, H. D. (1972). Information theory and the analysis of hospital cost structure. *Canadian Journal of Economics,* 5(3), 398–418.

Farley, D. E. (1989). Measuring casemix specialization and the concentration of diagnoses in hospitals using information theory. *Journal of Health Economics*, 8(2), 185–207.

Farley, D. E. and Hogan, C. (1990). Case-mix specialization in the market for hospital services. *Health Services Research,* 25(5), 757–783.

Gorina, Y, Hoyert, D., Lentzner, H. and Goulding, M. (2006). *Trends in causes of death among older persons in the United States.* Aging Trends, No 6. Hyattsville, Maryland: National Center for Health Statistics.

Herdan, G. (1957). The mathematical relation between the number of diseases and the number of patients in a community. *Journal of the Royal Statistical Society, Series A (General)*, 120(3), 320–330.

Hornbrook, M. C. (1982a). Hospital case mix: its definition, measurement and use. Part I. The conceptual framework. *Medical Care Review*, 39(1), 1–43.

Hornbrook, M. C. (1982b). Hospital case mix: its definition, measurement and use. Part II. Review of alternative measures. *Medical Care Review*, 39(2), 73–123.

Hurlbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology,* 52(4), 577–586.

Izsák, J. (1986). Measuring the secular changes of the concentration of death causes. *Genus*, 42(3–4), 197–208.

Izsák, J. (1989). Numerical experimentation on diversity curves applied in epidemiology. *Biométrie – Praximétrie,* 29, 97–113.

Izsák, J. (1993). Comparative analysis of death cause diversity curves in various countries. *Genus,* 49(1–2), 67–77.

Izsák, J. (2005). A pilot study on the frequency structure of histological neoplasm diagnoses in rats. *Journal of Theoretical Biology,* 236(4), 427–437.

Izsák, J. (2007). Parameter dependence of correlation between the Shannon index and members of parametric diversity index family. *Ecological Indicators,* 7(1), 181–194.

Izsák, J. and Juhász-Nagy, P. (1980–1981). Investigation on diversity of Hungarian mortality statistics. *Annales Universitatis Scientiarum de Rolando Eötvös Nominatae, Sectio Biologica,* 22–23, 35–45.

Izsák, J. and Juhász-Nagy, P. (1984). Studies of diversity indices on mortality statistics. *Annales Universitatis Scientiarum de Rolando Eötvös Nominatae, Sectio Biologica,* 24–26, 11–27.

Lasser, K. E., Himmelstein, D. U. and Woolhandler, S. (2006). Access to care, health status, and health disparities in the United States and Canada: results of a cross-national population-based survey. *American Journal of Public Health*, 96(7), 1300–1307.

McMonagle, C., Brown, D., Reeve, R. and Mancy, R. (2022). Trends in the diversity of mortality causes and age-standardised mortality rates among subpopulations within Scotland, 2001–2019. *SSM – Population Health,* 19, 101192.

Moriyama, I. M., Loy, R. M. and Robb-Smith, A. H. T. (2011). *History of the Statistical Classification of Diseases and Causes of Death.* Edited and updated by Rosenberg, H. M. and Hoyert, D. L. Hyattsville, MD: National Center for Health Statistics.

NCHS (2021). Mortality tables. *National Vital Statistics System, National Center for Health Statistics.* https://www.cdc.gov/nchs/nvss/mortality_tables.htm

Ostermann, T. and Schuster, R. (2015). An information-theoretical approach to classify hospitals with respect to their diagnostic diversity using Shannon's entropy. *Proceedings of the International Conference on Health Informatics*, 325–329.

Trias-Llimós, S. and Permanyer, I. (2023) Cause-of-death diversity from a multiple-cause perspective in the United States. *Demography*, 60(1), 73–98.