Journal of Clinical Epidemiology ■ (2015) ■

Journal of Clinical Epidemiology

ELSEVIER

ORIGINAL ARTICLE

Data on the distribution of cancer incidence and death across age and sex groups visualized using multilevel spie charts

Dror G. Feitelson*

Department of Computer Science and Engineering, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

Accepted 3 November 2015; Published online xxxx

Abstract

Cancer incidence and death statistics are typically recorded for multiple age and sex brackets, leading to large data tables which are difficult to digest. Effective visualizations of this data would allow practitioners, policy makers, and the general public to comprehend the data more readily and act on it appropriately. We introduce multilevel spie charts to create a combined visualization of cancer incidence and death statistics. Spie charts combine multiple pie charts, where the base pie chart (representing the general population) is used to set the angles of slices, and the superimposed ones use variable radii to portray the cancer data. Spie charts of cancer incidence and death statistics from Israel for 2009−2011 are used as an illustration. These charts clearly show various patterns of how cancer incidence and death distribute across age and sex groups, illustrating (1) absolute numbers and (2) rates per 100,000 population for different age and sex brackets. In addition, drawing separate charts for different cancer types illustrates relative mortality, both (3) across cancer types and (4) mortality relative to incidence. Naturally, this graphical depiction can be used for other diseases as well.

Keywords: Cancer incidence data; Mortality data; Age/sex distribution; Spie chart; Data visualization; Distribution comparison

1. Introduction

Humans have always been interested in health, sickness, and death. Health administrators in particular collect and use health and death statistics on a daily basis. Such statistics typically come in either of two forms. Professionals use multiple tables with hundreds of numbers, providing data about multiple factors in excruciating detail. The general public typically sees only artistic infographics, in which very few numbers are illustrated. There have been no effective visualizations of detailed data, despite the fact that graphic visualization of data contributes significantly to making it understandable and actionable, especially when large amounts of data are involved (e.g., [1,2]).

To appreciate the problem in the specific case of cancer statistics, consider the many different bisections and normalizations that are often used. Overall cancer rates for the entire population may be used for comparison of countries but are not very meaningful because they change considerably for different ages. There are also differences between the sexes. Data are therefore typically partitioned into multiple age brackets for either sex. These can then be normalized relative to an international standard population, thereby enabling valid international comparisons of real differences in cancer incidence and avoiding artifacts due to differences in population structure. Another choice is to present the data in absolute numbers—thereby placing the focus on the number of individuals affected—or in cases per 100,000 residents—thereby emphasizing hazard rates. In many cases, this is also done separately for different ethnic groups, to reflect differences in genetic background. Finally, the whole thing can be repeated for each type of cancer separately.

Spie charts are a statistical graphic based on pie charts which facilitates the comparison of two partitionings [3]. In particular, it has been suggested that this be used to portray hazards for different sex and age groups, such as the risk of being involved in a traffic accident. We extend this to multilayer spie charts and use them to portray both the cancer incidence rate and the cancer death rate for different sex and age combinations. In addition, both the relative absolute numbers of cases of different cancer types and the hazard rates per 100,000 residents are shown, using

Funding: This study was supported in part by an IBM Faculty Award.
Conflict of interest: None.
* Corresponding author. Tel./fax: +972-2-5494555.
E-mail address: feit@cs.huji.ac.il

http://dx.doi.org/10.1016/j.jclinepi.2015.11.001
0895-4356/

What is new?

- A graphical method to present detailed cancer incidence and death statistics in lieu of tables, thereby making the data more accessible and useful.
- Multilevel spie charts which combine both incidence and death data and relate them to the base population.
- An illustration of the methodology showing how patterns are identified.

slice areas and radii, respectively. It is expected that using such charts, it will be much easier to appreciate the patterns in cancer data at a glance.

2. Motivating example

As a motivating example, consider cancer incidence and death data for both sexes and multiple age brackets. An example of such data covering the years 2009−2011 from Israel is shown in Table 1. The data in this table are the average number of cases per year for each age and sex combination. All data are from the Israel Central Bureau of Statistics; a span of 3 years was used to reduce the effect of yearly fluctuations due to small numbers. If in addition we have a similarly sized table with the total population in each sex and age bracket, we can divide the cancer cases data by the population data to obtain hazard rates, that is the number of cases per 100,000 people in each sex and age group. The result of doing so is shown in Table 2.

Given such data, the typical way to visualize it would be some graph where the horizontal axis represents age, and the number of cases or the hazard rate is shown for both sexes as a function of age. However, it is not easy to come up with a graph that shows all the data. One problem is that the scales have different units: numbers of cases are not in the same scale as cases per 100,000 residents. Therefore, one has to compromise and show either number of cases or hazard rates, but not both. Wainer [4] suggests that in this case, a line plot can be rather effective, for example showing how the hazard rate changes with age for the two sexes, and even comparing the rate at each age bracket to the population-wide average by adding a horizontal line denoting the average. An example of such a graph is shown in Fig. 1, where both cancer incidence rates and cancer death rates are plotted for both sexes. Obviously, cancer incidence and death grow steeply with age, but deaths tend to occur about 10 years later. The patterns of cancer incidence are different for the two sexes: females are more susceptible at younger ages and males at older ages. In particular, women older than age 45 years already have a higher than average incidence rate, but for men, this happens only at age 55 years. As expected (and shown below), this is a result of the prevalence of breast cancer in women.

However, this is only half of the data, namely the hazard rates. So while we can see that the incidence and death rates are much higher for older people, we cannot tell how many cases are actually involved. To find the number of cases, we need to multiply the hazard rate by the population size in each age bracket and draw a separate line plot where the vertical axis represents cases rather than rates.

So can all the data be shown together? They can by using spie charts. The idea is simple: given that the number

Table 1. Cancer incidence and death data from Israel

	Incidence		Death	
Ages	**Males**	**Females**	**Males**	**Females**
0−14	177.00	138.67	31.00	22.33
15−24	172.33	194.67	27.33	18.00
25−34	314.67	841.67	41.33	45.67
35−44	488.67	1387.33	100.67	158.33
45−54	1171.33	2191.00	332.00	396.33
55−64	3138.67	3350.33	932.33	867.00
65−74	3567.33	3013.33	1231.67	1030.67
75+	3730.33	3706.33	2429.00	2533.33

Total cases averaged over the years 2009−2011.

Table 2. Cancer incidence and death hazard rates (cases per 100,000 individuals)

	Incidence		Death	
Ages	**Males**	**Females**	**Males**	**Females**
0−14	16.19	13.34	2.84	2.15
15−24	28.69	33.74	4.55	3.12
25−34	56.05	150.41	7.36	8.16
35−44	103.44	289.39	21.31	33.03
45−54	306.47	543.81	86.87	98.37
55−64	935.80	912.65	277.98	236.18
65−74	1977.46	1414.71	682.74	483.88
75+	2562.04	1736.80	1668.27	1187.13

This is obtained by dividing the data in Table 1 by the population of each sex and age bracket.

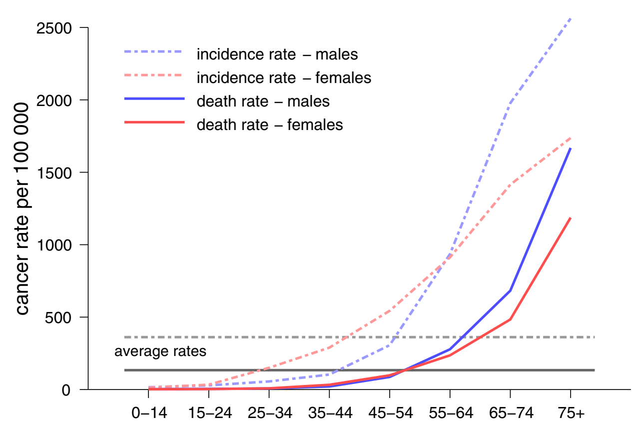

Fig. 1. Rendition of cancer incidence and death rates using line plots.