SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 89–98
doi:https://doi.org/10.37830/SJS.2022.1.07

OFFICIAL STATISTICS

Use of death statistics according to cause of death in health research

Gregorio Barrio

Biomedical Research Center Network for Epidemiology and Public Health (CIBERESP)
National School of Public Health, Carlos III Health Institute, gbarrio@isciii.es
Received: November 3, 2022. Accepted: December 26, 2022.

Abstract: Estimates of total and cause-specific mortality rates require information on the number of deaths (numerator) and the population at risk (denominator). In unlinked mortality studies, the numerator and denominator come from different sources, so there may be a numerator/denominator bias when estimating mortality rates according to certain individual attributes. This bias does not occur in linked mortality studies, in which data from the census or general population surveys are linked to vital records, and in the case of death, to the date and cause of death. However, regulations to protect individuals' confidentiality greatly limit the use of linked and unlinked mortality statistics for scientific research, whether due to the regulations themselves or because of the restrictive interpretations thereof by some statistical offices not always sufficiently argued. On the other hand, some methodological developments by these offices are of enormous relevance, for example, the linkage between socioeconomic indicators and mortality by the National Statistics Institute of Spain, which enables the study of the relationship between socioeconomic factors and mortality and its variation over time.

Keywords: Cause of death, Confidentiality, Health research, Mortality registers, Numerator-denominator bias, Record linkage

MSC: 60E05, 62P99

1 Introduction

The creation of civil registries in the 19th century to collect data on deaths and other sociodemographic characteristics was an important milestone for health research. The information contained in the death certificates, necessary for registering the deceased in these registries, began to be used in different studies on mortality according to factors such as occupation and place of residence. William Farr was the first to use these vital records to calculate the mortality rate across various occupations and in different geographical areas, in the mid-19th century in England and Wales

© INE

(Drever and Whitehead, 1997). The denominator for calculating this rate came from the population census. Farr passionately advocated for the development of a standard international nomenclature for collecting statistics on cause of death, coming to regard it as even more important to research than establishing a standard system of weights and measures in the physical sciences.

In the early 20th century, researchers in England and Wales also began to use information on occupation, as provided by the census and by death registries, to calculate mortality rates in different social strata. They used a social class scheme developed in 1911, which categorized occupations based on their social prestige. Known as the Registrar General's social class scheme, it was used throughout the 20th century in countless studies by British authors. Researchers from other countries have also developed similar classifications to study socioeconomic differences in mortality.

On the other hand, scientists in numerous countries have used census and other population-based data to characterize geographic or political-administrative areas based on demographic, socioeconomic, or environmental variables, using this information to assess the relationship between different geographic areas and the mortality of the population residing in them. On other occasions, investigators have evaluated the impact of an unexpected event or a health intervention on population health by comparing the mortality rates between areas or over time.

Thus, despite their limitations, cause-specific mortality statistics have helped define the main public health problems and carry out innumerable studies on the epidemiology and natural history of diseases. In fact, Hill considered that vital statistics laid the groundwork for the birth of epidemiology, and indeed, Snow used London's vital statistics, provided by the Registrar General in the mid-nineteenth century, in his landmark study of cholera transmission in that city (Hill et al., 1955).

2 Mortality studies with unlinked information

One of the earliest insights in occupational medicine was the recognition of the healthy worker effect (Checkoway et al., 2004). Occupational studies of mortality at the end of the 19th century described this bias after observing lower mortality in people who were employed relative to the general population. This phenomenon is due to the fact that people with a chronic disease are less likely to enter or remain in the labor market.

Likewise, for most of the 20th century, studies of socioeconomic inequalities in mortality have used the death registry and census data. In countries where these data were available (usually because occupation was recorded on the death certificate), researchers were able to describe the evolution of socioeconomic differences in overall and cause-specific mortality. For example, comparative studies in several European countries found that in the 1990s, socioeconomic differences in mortality were smallest in southern European countries like Italy and Spain (Mackenbach et al., 1997).

Since the 19th century, studies using a certain characteristic of the geographic or political-administrative area as the main independent variable have also generated knowledge of interest to public health. In the 19th century, rural populations had lower mortality than urban ones (Cosby et al., 2008) — a relationship that has been inverted in high-income countries. Other studies in these settings have investigated the relationship between various characteristics of the neighborhood of

residence and mortality, showing that the population residing in more deprived neighborhoods have higher mortality (Meijer et al., 2012).

Similarly, the availability of data on deaths and on populations in cities, municipalities, regions, and countries has made it possible to assess the impact of periods of high air pollution, heat waves, macroeconomic fluctuations, public health regulations, or access to medicines on overall and cause-specific mortality. For example, one study showed a sharp acceleration in the decline of mortality due to hepatitis C and other related causes, such as liver cancer and HIV infection, after the implementation of the Hepatitis C Strategic Plan in Spain, in April 2015, whose main component was providing universal and free access to direct-acting antivirals against this disease (Table 1) (Politi et al., 2022).

	Annual percent change in mortality rate(*)	
Cause of death	**Pre-intervention period**	**Post-intervention period**
Hepatitis C	-3.2	-18.4
Hepatocarcionoma	-0.9	-2.7
Cirrhosis	-3.7	-3.7
HIV disease	-8.3	-15.6
Non-C hepatitis	-5.8	-1.4
Other live diseases	-3.1	-1.6
All non-hepatic causes	-2.2	0.1

(*) 1. The pre-intervention period included from the first quarter 2001 to the first quarter 2015, inclusive. The post-intervention period included from the second quarter 2015 to the last quarter 2018.

Table 1: Comparison of mortality trends from hepatitis C and other hepatic and non-hepatic causes of death in the general population, before and after the implementation of the Strategic Plan for Tackling Hepatitis C: Spain, 2001 – 2018

Scientists' access to the data needed for such studies is uneven, which helps explain the absence of these types of investigations in some countries or regions. Restrictions are rooted in the fact that some statistical offices consider that removing the personal identification of the deceased —first name, last name, personal identification number— is not enough to protect individuals' confidentiality. If the population is small, they include characteristics like occupation, day of death, neighborhood, or municipality of residence within the scope of statistical confidentiality. There are even statistical offices that consider individual age as an object of special protection, so instead of providing the age of each deceased, they share only the five-year age group to which the deceased belongs. Others consider that it is the combination of variables that should remain secret, so they do not provide both age and cause of death for the deceased. There are endless ways of thinking on this matter.

Such obstacles make it difficult, if not impossible, to perform spatiotemporal analyses of great epidemiological and public health interest. For example, restrictions on data access preclude the study of mortality and daily ecological variables (e.g., temperature, air pollution) or socioeconomic indicators about the neighborhood or municipality of residence. It is also difficult to develop and

evaluate highly relevant public health interventions. In some cases, researchers can access these data once they formalize and fulfill certain administrative requirements imposed by the statistical offices. But many give up or do not even attempt it in the face of the heavy bureaucratic burden entailed.

Some scientists are surprised at this limitation, first of all, because individual observations are never disseminated in the findings of medical research, except in some clinical studies based on a very few patients. Second, the characteristics subject to special statistical protection are instrumental to generating results of interest. Third, the data that some offices treat as a statistical secret are not considered as such by others. These scientists probably forget that, in statistical offices, as in many other places (including research centers), ethics committees establish these limitations according to different criteria, either due to variations in national regulations on data protection or in the interpretation of these laws by different offices.

These heterogeneous interpretations can generate paradoxical situations, as in Spain, where for reasons of confidentiality, the National Statistics Institute does not routinely provide some attributes of the deceased person's microdata file, while some regional statistics offices do.

3 Mortality studies with linked information

In classic mortality studies by occupation or social class, researchers have calculated the number of deaths occurring among people of a given occupation during a given time period, divided by the number of people in that occupation for half the period. As noted, the data source for the numerator was the death certificate, and for the denominator the population census. As the numerator and denominator come from different sources, these are unlinked cross-sectional studies, which are at risk of a numerator/denominator bias, since a person's occupation in the death registry may not match that person's occupation in the census (Lynge, 2011).

This bias can also occur with other variables, such as sex or age, since both come from different sources; some people may appear as men in the death registry and as women in the population census, or vice versa, and the age may different. However, the scientific literature has never made any reference to these errors because they are probably infrequent and have a negligible impact on the study results.

Starting in the second half of the 20th century, the central statistics offices of several countries began to link data from the census and the death registry in individual entries, making it possible to avoid this numerator/denominator bias in research. The central statistics offices provided researchers with the linked data set, with census variables along with the date and cause of death, after removing personal identifiers to protect privacy. The first countries to implement this methodology were the USA, in the 1960s, followed by Denmark, Finland, Norway, England, Wales, and France in the 1970s. Some countries established this linkage in the entire population, while others did so only with a representative census sample (Fox, 1989). Subsequently, some central statistics offices began to link data from general population surveys to that from the death registry (Duleep, 1989). Since then, and especially from the 1990s, the statistical offices of other countries have followed suit, implementing the methodologies needed to create a linked data set. In Spain and Italy, some regional statistics offices have been ahead of the national statistical offices in that regard. In Spain, this change was of great importance due to the decrease in the proportion of

death certificates containing the occupation of the deceased, which made it impossible to perform mortality studies according to this variable. The creation of a linked data set enabled researchers to continue investigating the relationship between socioeconomic status and mortality.

Population and educational level(*)	Nineties	First half of' the 2000s'	Nineties	First half of the 2000s'
Finland				
Low	1.97	2,08	1,59	1,84
Medium	1.6	1.61	1.22	1.35
High	1	1	1	1
Sweden				
Low	1.78	1.9	1.88	1.88
Medium	1.42	1.47	1.47	1.42
High	1	1	1	1
Norway				
Low	1.88	2.35	1.76	2.12
Medium	1.43	1.63	1.3	1.45
High	1	1	1	1
Denmark				
Low	1.77	2	1.62	1.85
Medium	1.46	1.53	1.24	1.34
High	1	1	1	1
France				
Low	2.23	2.37	1.64	1.8
Medium	1.6	1.7	1.17	1.38
High	1	1	1	1
Switzerland				
Low	1.95	2.22	1.43	1.54
Medium	1.41	1.52	1.11	1.13
High	1	1	1	1
Region of Madrid				
Low	1.55	1.56	1.37	1.3
Medium	1.3	1.27	1.18	1.23
High	1	1	1	1
Basque Country				
Low	1.49	1.51	1.25	1.39
Medium	1.2	1.16	1.12	1.22
High	1	1	1	1

(*) Low level: lower than primary studies, primary studies and the first cycle of secondary education. Intermediate level: second cycle of secondary education and studies after secondary education. High level: university studies.

Table 2: Mortality rate ratio according to educational level in various European populations. Subjects from 30 to 74 years. Nineties and first half of the two thousand years

Comparative studies with this type of data from various Western European countries have confirmed the smaller socioeconomic differences in mortality in southern compared to northwestern European countries (Mackenbach et al., 2015). These results are consistent regardless of whether

the measure of socioeconomic status is based on occupation or educational attainment. Table 2 shows these findings according to education. Data for Spain and Italy are generally collected and reported at a regional level, but this pattern is similar in analyses of data in the entire population.

Apart from avoiding the numerator/denominator bias, linked data sets are fertile grounds for testing hypotheses because of the large number of variables they contain. For example, based on a linkage between the 2001 population census and the death registry over the following ten years, implemented by the National Statistics Institute in Spain, a study found an acceleration in mortality decline during the 2008 economic crisis with respect to the previous period, which was more pronounced in people with low compared to high socioeconomic status (Table 3). In that study, the size of the dwelling (m^2) and the number of cars in the household, as recorded in the 2001 population census, were used as indicators of socioeconomic status (Regidor et al., 2016).

	APC in mortality rate		
Indicators of wealth	**(1) Precrisis (2004-2007)**	**(2) Crisis (2008-2011)**	**Effect size (2)-(1)**
Household floor space(m^2)			
Low (<72)	-1.7	-3.0	-1.4
Medium (72-104)	-1.7	-2.8	-1.1
High (>104)	-2.0	-2.1	-0.1
Household car ownership			
Low (no car)	-0.3	-2.3	-2.0
Medium (1 car)	-1.6	-2.4	-0.8
High (2 or more cars)	-2.2	-2.5	-0.3

Table 3: Table 3. Trends in premature mortality (lower than 75 years old) in Spain. Annual percent change (APC) in mortality rate before and during the 2008 economic crisis, and effect size according to indicators of wealth

Statistics offices apply the same interpretation of statistical confidentiality to linked and unlinked data, withholding information on variables they consider should not be disclosed. Furthermore, when population samples are linked to the death registry, statistics offices do not provide detailed information on some variables in the resulting data set. For example, they do not release the specific cause of death (at the level of the fourth digit of the International Classification of Diseases), but only the large disease group to which those causes belong.

Statistics offices consider that a low number of deaths from a specific cause cannot be used for statistical analysis. This paternalistic criterion is inadequate, since it is impossible to know the hundreds of possibilities that the availability of the specific cause of death would provide in answer to an infinite number of research questions. Statistics offices have absolute legitimacy when determining what data are subject to statistical confidentiality, based on whatever law, moral reasoning, or ethical criteria they deem appropriate to apply. But they lack legitimacy when they resort to supposedly technical criteria without adequate theoretical and empirical arguments to support them.

A common feature of linked mortality studies in many countries is that investigators often have little or no control over the linkage processes or the techniques used to build the database. To make

matters worse, they may also have little information about the processes and techniques used. This knowledge is essential because linkage errors, materialized in the impossibility of linking some individual records (missing links) or falsely linking records, could generate bias when they do not occur randomly. In fact, scientific journals often ask researchers for details about the process before publishing the results. It is understandable that statistics offices prevent access to individualized records with personal identifiers, but it is highly questionable that they rarely provide basic information on linkage techniques or the validity of the linkage made. It is also problematic that researchers do not participate in the planning of these processes, since it could allow them to assess the quality of the linkage, reducing errors and enabling a better interpretation of the study results drawn from these linked databases (Harron et al., 2017; Harron, 2022).

4 Other mortality studies with linked information

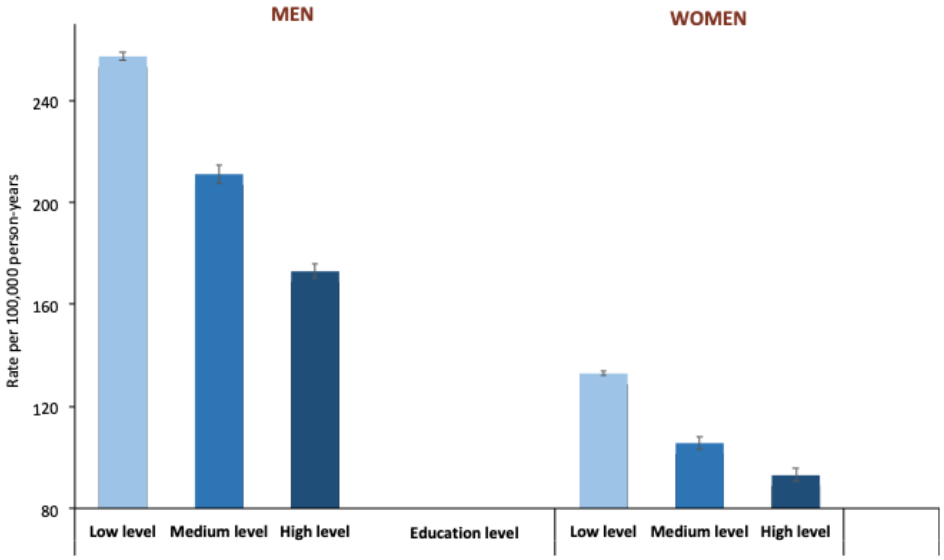
Another methodological option that avoids numerator/denominator bias in studies of socioeconomic characteristics and mortality is that developed by the National Statistics Institute of Spain in the continuous updating of population figures. By crossing numerous sources of administrative information, all the socioeconomic indicators collected in any of those sources are made available for each citizen. Subsequently, the National Statistics Institute links these variables to each deceased person in the death registry, thus avoiding the numerator/denominator bias when calculating mortality rates based on these attributes.

These population and mortality data, plus data on sex, age, and level of studies, were used to calculate the estimates that appear in Figure 1. In mortality from causes of death strongly related to alcohol, there is an inverse gradient according to the level of studies in both men and women, a result that is similar to those from other high-income countries. Likewise, in Spain and other countries, there is an inverse gradient in alcohol intake according to the level of studies in men, while in women, alcohol intake is most frequent in those with a high level of studies (Boyd et al., 2021; Donat et al., 2022). This discordance between the findings on alcohol-related mortality and alcohol intake in women has been called the alcohol harm paradox, the reasons for which are still unclear among the international scientific community.

5 Epidemiological follow-up studies

In epidemiological follow-up studies, a large amount of information is obtained over time from a large sample of research subjects. The data collection questionnaires include numerous variables obtained through personal interviews, physical tests, and blood and urine analyses. In this way it is possible to estimate the relationship between a wide variety of factors and the appearance of diseases and cause-specific mortality.

Researchers turn to statistics offices to find out the vital status of research subjects and, if deceased, the cause of death. Some offices make this information contingent on different agreements and protocols with researchers, so that investigators can obtain the information they need from the death registry. In Spain, to determine an individuals's vital status, researchers can request the data from either the National Statistics Institute or from the National Death Index, a database



1.Alcohol-induced pseudo-Cushing's syndrome (E24.4), alcohol use disorders (F10), alcoholic nervous system degeneration (G31.2), alcoholic polyneuropathy (G62.1), alcoholic myopathy (G72.1), alcoholic cardiomyopathy (I42.6), alcoholic gastritis (K29.2), alcoholic liver disease (K70), alcohol-induced acute pancreatitis (K85.2), alcohol-induced chronic pancreatitis (K86.0), maternal complication of foetal alcohol injury (O35. 4), foetus and newborn affected by maternal alcoholism (P04.3), foetal alcohol dysmorphic syndrome (Q86.0), blood alcohol finding (R78.0), accidental alcohol exposure poisoning (X45), intentional self-inflicted alcohol exposure poisoning (X65), alcohol exposure poisoning, undetermined intent (Y15) and evidence of alcohol involvement (Y90-Y91).Chronic hepatitis, not elsewhere classified (K73) and fibrosis and cirrhosis of liver (K74, except biliary cirrhosis -K74.4 to K74.5-). Cancers of lip, oral cavity and pharynx (C00-C13), oesophagus (C15), larynx (C32) and liver (C22). Tuberculosis (A15-A19, B90, K67.3, P37.0) and lower respiratory infection/pneumonia (A48.1, A70, J09-J15.8, J16, J20-J21, P23.0-P23.4), pancreatitis (K85-K86, except alcohol-induced pancreatitis ?K85.2 and K86.0-) and epilepsy (G40-G41).

Figure 1: Age-standardized mortality rate from causes closely related to alcohol in people aged 25 years, by educational attainment. Spain, 2016-2019

managed by the Ministry of Health; however, this index does not provide access to the cause of death.

Based on the results obtained in some of these investigations, it is possible to quantify the impact that certain circumstances or factors may have on the burden of disease and death in the population. For example, using estimates on the relationship that tobacco and alcohol consumption have with mortality from various causes of death, together with information on the prevalence of these behaviors, it is possible to estimate deaths potentially attributable to tobacco and alcohol from various causes of death, together with information on the prevalence of these behaviors, it is possible to estimate deaths potentially attributable to tobacco and alcohol. According to two studies, 14.0 and 4.0 of deaths were attributable to tobacco and alcohol use, respectively, over the first two decades of the 21st century in Spain (Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI), 2016; Donat et al., 2022).

6 Conclusions

Mortality registries and cause-of-death statistics are of immense importance to clinical and public health research, far outweighing the value of existing morbidity records. In Spain, for example, the quality of the mortality data and their value have increased considerably by including new sociodemographic variables, such as educational level or occupation. In addition, the validity of registered causes of death has increased by better incorporating judicial and forensic information. However, fully capitalizing on the potential of these data is still not possible due to limitations in accessing some variables, such as the day or municipality of death. These restrictions derive from a very conservative interpretation of personal data protections by the ethics committees in some statistics offices. This rigid position should be reconsidered, and efforts made to design simple procedures so that scientists can access this information —with privacy guarantees but without tedious bureaucratic procedures. After all, the greatest benefits for the population derive, surely, from achieving an adequate balance between protecting people's right to privacy and carrying out research that improves their health and quality of life.

On the other hand, the usefulness of mortality registries for health research would be greatly improved if investigators requesting linkages to other registries could participate in some way in planning the linking procedures or at least receive detailed information about them. In this way, they could more adequately interpret the findings of their research and respond with confidence to the editors of the journals that disseminate their work.

References

Boyd, Jennifer, Clare Bambra, Robin C Purshouse, and John Holmes (2021). Beyond behaviour: How health inequality theory can enhance our understanding of the 'alcohol-harm paradox'. *International Journal of Environmental Research and Public Health 18*(11), 6025.

Checkoway, Harvey, Neil Pearce, and David Kriebel (2004). *Research methods in occupational epidemiology*, Volume 34. Monographs in Epidemiology.

Cosby, Arthur G, Tonya T Neaves, Ronald E Cossman, Jeralynn S Cossman, Wesley L James, Neal Feierabend, David M Mirvis, Carol A Jones, and Tracey Farrigan (2008). Preliminary evidence for an emerging nonmetropolitan mortality penalty in the united states. *American Journal of Public Health 98*(8), 1470–1472.

Donat, Marta, Gregorio Barrio, Juan-Miguel Guerras, Lidia Herrero, José Pulido, María-José Belza, and Enrique Regidor (2022). Educational gradients in drinking amount and heavy episodic drinking among working-age men and women in spain. *International Journal of Environmental Research and Public Health 19*(7), 4371.

Drever, Frances and Margaret Whitehead (1997). *Health inequalities: decennial supplement*. Decennial supplement. Series DS No. 15. London: The Stationery Office, London.

Duleep, Harriet Orcutt (1989). Measuring socioeconomic mortality differentials over time. *Demography 26*(2), 345–351.

Fox, AJ (1989). Longitudinal studies based on vital registration records. *Revue D'épidémiologie et de Santé Publique 37*(5-6), 443–448.

Harron, Katie (2022). Data linkage in medical research. *BMJ Medicine 1*(1).

Harron, Katie L, James C Doidge, Hannah E Knight, Ruth E Gilbert, Harvey Goldstein, David A Cromwell, and Jan H van der Meulen (2017). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology 46*(5), 1699–1710.

Hill, AB et al. (1955). Snow-an appreciation. *Proceedings of the Royal Society of Medicine 48*(12), 1008–12.

Lynge, Elsebeth (2011). Occupational mortality. *Scandinavian Journal of Public Health 39*(7_suppl), 153–157.

Mackenbach, Johan P, Ivana Kulhánová, Gwenn Menvielle, Matthias Bopp, Carme Borrell, Giuseppe Costa, Patrick Deboosere, Santiago Esnaola, Ramune Kalediene, Katalin Kovacs, et al. (2015). Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 european countries. *Journal of Epidemiology and Community Health 69*(3), 207–217.

Mackenbach, Johan P, Anton E Kunst, Adriënne EJM Cavelaars, Feikje Groenhof, and Jose JM Geurts (1997). Socioeconomic inequalities in morbidity and mortality in western europe. *The lancet 349*(9066), 1655–1659.

Meijer, Mathias, Jeannette Röhl, Kim Bloomfield, and Ulrike Grittner (2012). Do neighborhoods affect individual mortality? a systematic review and meta-analysis of multilevel studies. *Social Science & Medicine 74*(8), 1204–1212.

Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI) (2016). Muertes atribuibles al consumo de tabaco en españa, 2000-2014. Technical report, Ministerio de Sanidad, Servicios Sociales e Igualdad, Madrid.

Politi, Julieta, Juan-Miguel Guerras, Marta Donat, María J Belza, Elena Ronda, Gregorio Barrio, and Enrique Regidor (2022). Favorable impact in hepatitis c–related mortality following free access to direct-acting antivirals in spain. *Hepatology 75*(5), 1247–1256.

Regidor, Enrique, Fernando Vallejo, José A Tapia Granados, Francisco J Viciana-Fernández, Luis de la Fuente, and Gregorio Barrio (2016). Mortality decrease according to socioeconomic groups during the economic crisis in spain: a cohort study of 36 million people. *The Lancet 388*(10060), 2642–2652.