Estimating bias in educational inequalities in mortality: A simulation study

Ana C. Gómez-Ugarte[1], Ugofilippo Basellini[1], Carlo G. Camarda[2], Fanny Janssen[3,4], and Emilio Zagheni[1]

[1] *Max Planck Institute for Demographic Research, Rostock, Germany*
[2] *Institut national d'études démographiques, Aubervilliers, France*
[3] *Aging and Longevity, Netherlands Interdisciplinary Demographic Institute - KNAW/University of Groningen, The Hage, The Netherlands*
[4] *Population Research Centre, Faculty of Spatial Sciences, University of Groningen, Groningen, The Netherlands*

September 11, 2024

Abstract

In many countries, unlinked cross-sectional data is the only available data to study educational inequalities in mortality. However, such data is subject to three data-quality issues: under-coverage, age misreporting and education misreporting. Many studies have looked into the first two issues, whereas education misreporting has been less explored. The first goal of this study is to provide a comprehensive evaluation of how estimates of educational inequalities in mortality can be affected by this source of error. For this, we rely on simulation scenarios with varying direction and magnitude of education misreporting as well as on different measures of educational inequalities. We find that education overstatement downward biases educational inequalities in mortality based on life expectancy and upward biases those based on life span variation measures, while education understatement upward biases inequality. The second goal of this paper is to quantify the degree of education misreporting in a given dataset, and to introduce an adjustment procedure to correct educational misreporting. Previous studies have shown that education-specific mortality rates converge at older ages and that crossovers should only happen at very old ages. Starting from these theoretical framework, we adjust the distance between education-specific log-mortality rates such that its pattern follows the theoretical and observed patterns. This, in turn, allows us to adjust the education-specific log-mortality rates and derive a new estimate of educational inequalities in mortality.

Introduction

In many countries, unlinked cross-sectional data is the only available data to study educational differentials in mortality. However, such data is subject to three primary data-quality issues: under-coverage, age misreporting and education misreporting. Many studies have looked into the first two issues (Hill et al., 2009; Palloni et al., 2021; Schmertmann et al., 2024). Conversely, education misreporting has been less explored as it can only be

assessed in a few countries where both linked and unlinked data sources are available. The few studies that focus on this problem are empirical analysis that focus on a specific country (Rostron et al., 2010; Sorlie and Johnson, 1996), making it difficult to generalise and apply elsewhere. Moreover, previous studies show that this bias is not constant over time, that it changes by country and that it affects people on either end of the education distribution (Jasilionis and Leinsalu, 2021; Shkolnikov et al., 2007).

Methods to assess the validity of unlinked cross-sectional data are scarce and mostly depend on visual checks to detect unusual age-patterns or mortality cross-overs (Mackenbach et al., 2015). In contexts with less ideal data quality, researchers must rely on data quality checks that perform comparisons with developed countries' patterns, although there is no clear knowledge whether less developed countries will follow the same trends. Solutions to try to limit the error of the estimates include limiting the age-range (Shkolnikov et al., 2022) or using modelled-data in problematic age-groups (Mackenbach et al., 2015), all of which are subjective as no objective criteria to guide researchers exists. Some studies have tried to control for these known errors by imputing missing data with conservative assumptions, and by excluding small minority groups known to be greatly affected by misreporting (Sasson, 2016).

When the only available data is the one that contains errors and no additional information exists, it is difficult to identify the magnitude or the direction of the error and an almost impossible task to recover the true values. A simulation study allows us to compare the outcomes with the true values instead of comparing them with the observed ones, which are subject to misclassification. For this reason, we rely on simulation scenarios with varying direction and magnitude to understand how education misreporting affects mortality estimates. Additionally, we evaluate how this in turn may bias estimates of educational inequalities in mortality.

As a second goal, we quantify the degree of education misreporting in a given dataset and introduce an adjustment procedure to correct educational misreporting. Starting from observations and a theoretical framework derived from previous studies, we adjust the distance between education-specific log-mortality rates such that its pattern follows the theoretical and observed patterns. This, in turn, allows us to adjust the education-specific log-mortality rates and derive a new estimate of educational inequalities in mortality.

Data and methods

Suppose we have n age groups denoted by $a = 1, \ldots, \omega$, and g number of education groups. Let $\boldsymbol{\gamma}^k = (\gamma_1, ..., \gamma_\omega)'$ be a vector of length n containing the true age-specific mortality rates of the k-th group, and $\boldsymbol{\delta}^k = (\delta_1, ..., \delta_\omega)'$ be the vector of true deaths by age. Then we can define a series of matrices by which we can introduce errors in the true mortality rates. We start by assuming that the errors are limited to the death counts, with the exposure being error free.

Let $c_x \in [0, 1]$ be the probability that the death of someone aged x years is registered, and let $\boldsymbol{c}^k = (c_1, ..., c_\omega)'$ denote the vector collecting all age elements for group k. We further expand the vector in an $n \times n$ diagonal matrix $\boldsymbol{C}^k = \text{diag}(\boldsymbol{c}^k)$.

Further, let $\boldsymbol{P}^k$ be the $n \times n$ age-misreporting matrix of the k-th group, with elements

p_{yx} denoting the probability that an individual of true age x reports it as y, where for all ages x it holds $\sum_i p_{ix} = 1$.

Moreover, let $\boldsymbol{E}^{ks}$ be the $n \times n$ matrix containing the education misreporting from group k to group s, with elements e_x^{ks}, denoting the probability that an individual of reported age x with true education k reports it as s. Here $\sum_j e_x^{js} = 1$ for all ages x.

Assuming we have 3 education groups, that is $g = 3$, which we denote as l, m and h, then the observed deaths in each education group are given by:

$$\begin{bmatrix} D^l \\ D^m \\ D^h \end{bmatrix} = \begin{bmatrix} E^{ll} & E^{ml} & E^{hl} \\ E^{lm} & E^{mm} & E^{hm} \\ E^{lh} & E^{mh} & E^{hh} \end{bmatrix} \begin{bmatrix} P^l & 0 & 0 \\ 0 & P^m & 0 \\ 0 & 0 & P^h \end{bmatrix} \begin{bmatrix} C^l & 0 & 0 \\ 0 & C^m & 0 \\ 0 & 0 & C^h \end{bmatrix} \begin{bmatrix} \delta_l \\ \delta_m \\ \delta_h \end{bmatrix} \tag{1}$$

If we let $\boldsymbol{n}^k = (n_1, ..., n_\omega)'$ denote the vector collecting population exposures of group k by age. Then, the vector of observed education specific mortality rates of the k-th education group is:

$$\mu^k = \left(\frac{D_\alpha^k}{n_\alpha^k}, ..., \frac{D_w^k}{n_w^k} \right)$$

Using the above formulation, we say that there are no errors in the observed deaths when $\mathbf{C}^k = \mathbf{P}^k = \mathbf{E}^{kk} = I$. Contrarily, the most common errors occur when:

- $\mathbf{C}^k \neq I$ (coverage errors).
- $\mathbf{P}^k \neq I$ (age-misreporting errors).
- $\mathbf{E}^{kk} \neq I$ (education-misreporting errors).

Therefore, we can evaluate how the mortality rates change when we introduce such errors. Under-registration and age-misreporting, for a single education group, have been deeply analyzed by Schmertmann et al. (2024). Here we focus on the education-misreporting errors and its interaction with the other two.

Data generating process

The example below is based on the mortality rates by education of Swedish females in 2016 estimated with data from Eurostat (2024a,b). The true death rates were smoothed using a generalized linear smoothing procedure.

Measures of socioeconomic inequalities in mortality

We estimate educational inequalities in mortality using a variety of measures that are often used in the literature (for a recent overview of their advantages and limitations, see Gómez Ugarte Valerio et al., 2024): range in life expectancy, ratio in life expectancy, range in standard deviation of the ages at death, ratio of the standard deviation of the ages at death, and population non-overlap index.

Measuring bias

Let θ and $\hat{\theta}$ be the value of the inequality measure estimated from the true and the observed mortality rates, respectively. Then, we define the relative bias as: $\frac{\hat{\theta}-\theta}{\theta}$.

Misreporting scenarios

To understand the impact of different data quality errors on education-specific mortality estimates, we define different data error scenarios by varying the direction and magnitude of the reporting errors. We start with the case of two education groups. Table 1 shows the assumptions made in each scenario. The first three scenarios assume constant misreporting rates across ages, while the following scenarios introduce differential education misreporting rates by age.

Scenario	Education misreporting	Age misreporting	Coverage
S1	X% of the low educated report high education Constant across ages	None	None
S2	X% of the high educated report low education Constant across ages	None	None
S3	X% of the low educated report high education and X% of the high educated report low education Constant across ages	None	None
S4	X% of the low educated report high education, the percentage increases linearly over age from 0% to X% in the last age group	None	None
S5	X% of the high educated report low education, the percentage increases linearly over age from 0% to X% in the last age group	None	None
S6	X% of the low educated report high education and X% of the high educated report low education, the percentage increases linearly over age from 0% to X% in the last age group	None	None

Table 1: Education misreporting scenarios for two education groups.

Preliminary results

Figure 1 shows how different education misreporting scenarios (1, 2 and 3 from Table 1) affect the Swedish females education-specific mortality rates.

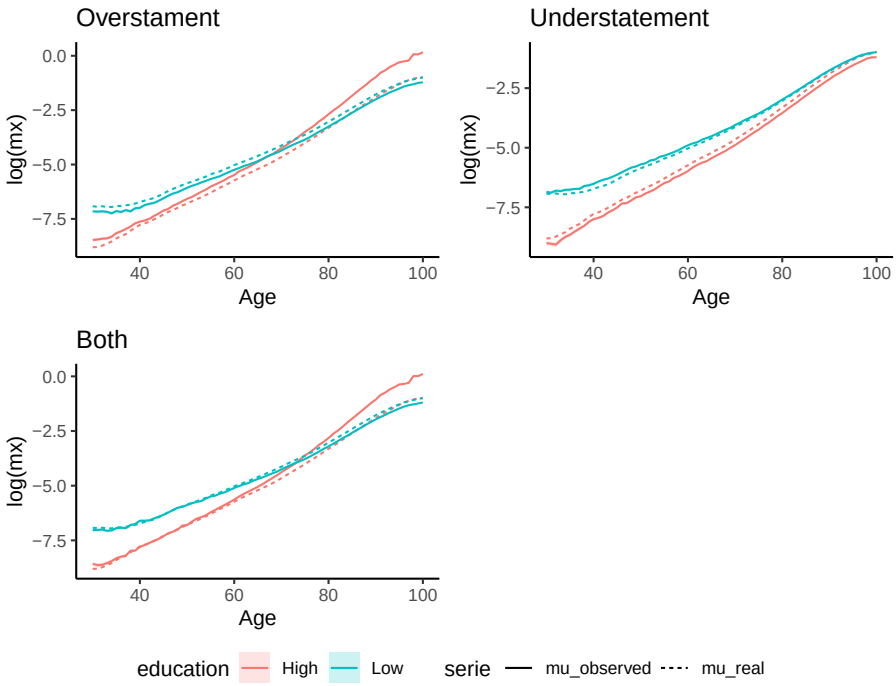

Figure 1: Education specific death rates under scenarios 1, 2 and 3 from Table 1 with 20% of records misreporting the respective education level. The dashed line represents the true mortality rate and the solid line the observed mortality rate.

Figure 2 shows the relative bias under different scenarios and degrees of education misreporting in several measures of educational inequality. Overall, educational overstatement causes a negative bias in measures based on life expectancy and a positive bias in measures of lifespan variation. Education understatement has the opposite effect. It seems the population non-overlap index increases regardless of the case.

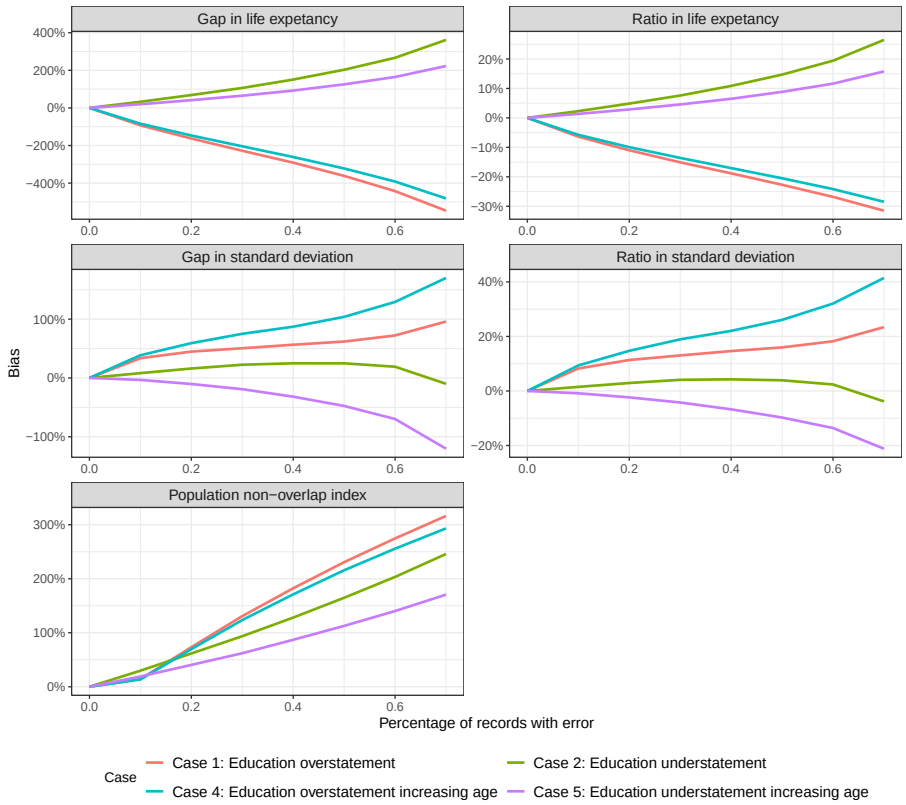

Figure 2: Relative bias in various inequality measures under scenarios 1, 2, 4 and 5 from Table 1.

Next steps

Next steps include extending the simulation scenarios tor three education groups and incorporating under-coverage of deaths and age-misreporting errors to understand how they interact with education misreporting.

In further steps, we will try to quantify the degree of education misreporting in a given dataset, and to introduce an adjustment procedure to correct educational misreporting in observed data when no additional information is available.

References

Eurostat (2024a). Deaths by age, sex and educational attainment level. Available at https://ec.europa.eu/eurostat/databrowser/product/view/demo_maeduc?lang=en.

Eurostat (2024b). Population on 1 January by age, sex and educational attainment level. Available at https://ec.europa.eu/eurostat/databrowser/product/view/demo_pjanedu?lang=en.

Gómez Ugarte Valerio, A. C., Basellini, U., Camarda, C. G., Janssen, F., and Zagheni, E. (2024). Reassessing socioeconomic inequalities in mortality via distributional similarities. Technical Report WP-2024-007, Max Planck Institute for Demographic Research, Rostock.

Hill, K., You, D., and Choi, Y. (2009). Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data errors. *Demographic Research*, 21:235–254.

Jasilionis, D. and Leinsalu, M. (2021). Changing effect of the numerator–denominator bias in unlinked data on mortality differentials by education: evidence from estonia, 2000–2015. *J Epidemiol Community Health*, 75(1):88–91.

Mackenbach, J., Menvielle, G., Jasilionis, D., and de Gelder, R. (2015). Measuring Educational Inequalities in Mortality Statistics. OECD Statistics Working Papers 2015/08.

Palloni, A., Beltrán-Sánchez, H., and Pinto, G. (2021). Estimation of older-adult mortality from information distorted by systematic age misreporting. *Population Studies*, 75(3):403–420.

Rostron, B. L., Boies, J. L., and Arias, E. (2010). *Education reporting and classification on death certificates in the United States*. U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Md., Washington, DC.

Sasson, I. (2016). Trends in Life Expectancy and Lifespan Variation by Educational Attainment: United States, 1990–2010. *Demography*, 53(2):269–293.

Schmertmann, C., Lanza Queiroz, B., and Gonzaga, M. (2024). Data errors in mortality estimation: Formal demographic analysis of under-registration, under-enumeration, and age misreporting. *Demographic Research*, 51:229–266.

Shkolnikov, V. M., Andreev, E. M., and Jasilionis, D. (2022). Changes in mortality disparities by education in Russia from 1998 to 2017: evidence from indirect estimation. *European Journal of Public Health*, 32(1):21–23.

Shkolnikov, V. M., Jasilionis, D., Andreev, E. M., Jdanov, D. A., Stankuniene, V., and Ambrozaitiene, D. (2007). Linked versus unlinked estimates of mortality and length of life by education and marital status: Evidence from the first record linkage study in lithuania. *Social Science Medicine*, 64(7):1392–1406.

Sorlie, P. D. and Johnson, N. J. (1996). Validity of education information on the death certificate:. *Epidemiology*, 7(4):437–439.