DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 51, ARTICLE 9, PAGES 229–266
PUBLISHED 6 AUGUST 2024
https://www.demographic-research.org/Volumes/Vol51/9/
DOI: 10.4054/DemRes.2024.51.9

Research Article

Data errors in mortality estimation: Formal demographic analysis of under-registration, under-enumeration, and age misreporting

Carl P. Schmertmann

Bernardo L. Queiroz

Marcos R. Gonzaga

Contents

Demographic Research: Volume 51, Article 9
Research Article

Data errors in mortality estimation: Formal demographic analysis of under-registration, under-enumeration, and age misreporting

Carl P. Schmertmann[1]

Bernardo L. Queiroz[2]

Marcos R. Gonzaga[3]

Abstract

BACKGROUND
Omissions and misreported ages in both death and exposure data cause bias in mortality and life expectancy estimates. Most discussions of data errors have focused on a single type of error only, and most rely on empirical examples rather than formal analysis.

OBJECTIVE
We wish to analyze data errors and their interactions in a single, coherent framework in which all three of the major data problems – death under-registration, census under-enumeration, and age misreporting – coexist and interact.

METHODS
We build a framework for decomposing the biases caused by various data errors in mortality rates and life expectancy calculations. In addition to purely mathematical analysis, we apply the calculations to mortality and population data from Brazil, a country with intermediate data quality.

CONCLUSIONS
Analytical and empirical calculations show that biases caused by data errors vary considerably across ages; that age misreporting has very small effects on life expectancy calculations at old ages; and that enumeration and registration errors are likely to cause much larger biases than age misreporting.

[1] Center for Demography and Population Health, Florida State University, Tallahassee, Florida, USA. Email: schmertmann@fsu.edu
[2] CEDEPLAR, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[3] PPGDEM, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil.

CONTRIBUTION
Combining an explicit analytical structure with empirical examples allows improved understanding of the consequences of data errors for mortality estimates in a wide variety of settings. It also provides insights for further study.

1. Introduction

Errors in death and exposure data cause bias in estimated mortality rates and life expectancy (Coale and Kisker 1990; Coale and Li 1991; Condran, Himes, and Preston 1991; Preston, Elo, and Stewart 1999; Jdanov et al. 2008; Palloni, Beltrán-Sánchez, and Pinto 2021). When death counts are less complete than population counts, for example, standard formulas will underestimate mortality rates and overestimate life expectancy. Similarly, the possible bias caused by age misreporting has been a central theme in the literature on mortality 'crossovers' (Coale and Kisker 1986; Nam 1995; Preston et al. 1996; Preston, Elo, and Stewart 1999; Di Lego, Turra, and Cesar 2017; Queiroz et al. 2020).

Most discussions of the effects of data errors on mortality estimates focus on a single type of error only (for example, under-registration of deaths or misreporting of ages). Most rely on empirical examples rather than formal analysis. Here we undertake a mathematical analysis of situations in which all three of the major data problems – death under-registration, census under-enumeration, and age misreporting – coexist and interact. We build a framework for decomposing the resulting errors in mortality rates and life expectancy calculations.

In addition to purely mathematical analysis, we apply our framework to an example population from a country with intermediate data quality (Brazil). Combining an explicit analytical structure with empirical examples allows improved understanding of the consequences of data errors for mortality estimates in a wide variety of settings. It also provides insights for further study.

2. Previous studies

Demographers have developed a variety of indicators to detect and measure the different types of errors that affect mortality estimates (Coale and Li 1991; Kannisto, Jeune, and Vaupel 1999; Jdanov et al. 2008; Palloni, Beltrán-Sánchez, and Pinto 2021). Death distribution methods, for example, are widely used to estimate completeness of death records relative to census enumeration (Bennett and Horiuchi 1981; Hill 1987; Hill, You, and Choi 2009). The literature on age misreporting includes methods to detect and evaluate digit preference, to estimate the impact of age errors on mortality levels and life

expectancy, to evaluate the quality of data about the number of centenarians, and to investigate the consequence of age errors for estimating mortality differentials between subgroups (Whipple 1919; Myers 1940; Bhat 1990; Pullum 1991; Kannisto, Jeune, and Vaupel 1999; Spoorenberg and Dutreuilh 2007; Gomes and Turra 2009; Nepomuceno and Turra 2020).

Empirical analyses of age misreporting focus on two main problems: age heaping and age overstatement among the very old. Whipple (1919) develops a pioneering method to detect heaping in demographic data disaggregated by single year of age. Age overstatement could cause dubious or irregular patterns of mortality rates at older ages, and has therefore been the subject of several studies (e.g., Coale and Li 1991; Kannisto, Jeune, and Vaupel 1999; Jdanov et al. 2008). Addressing both of the principal age reporting problems, Jdanov et al. (2008) proposes a classification system for national data quality, with four levels (best, acceptable, conditionally acceptable, and weak) that flag irregular mortality patterns at older ages.

Many studies have considered the consequences of differential age misreporting for comparisons of mortality between populations, especially between advantaged and disadvantaged subpopulations in developed countries (e.g., Nam 1995; Preston et al. 1996). Preston et al.'s (1996) classic study looks at misreporting's impact on levels of African American mortality at older ages.

The quality of age data is an even more central problem in less developed countries, and there are a number of valuable case studies. Dechter and Preston (1991) document age misreporting in Latin American data. Beltrán-Sánchez et al. (2020) and Palloni, Beltrán-Sánchez, and Pinto (2021) estimate levels of death counts under-registration for a series of Latin American countries and document issues of age misreporting for more recent periods in the region. There are also country-specific studies. In Brazil, for example, Martins (2022) tests alternatives for adjusting the age distribution of deaths to measure the impact of such errors in life expectancy, and Turra et al. (2023) examines the quality of age declarations by comparing COVID-19 vaccination records to other sources. In South Africa, Richman (2017) demonstrates that age overstatement in both population and death counts biases mortality comparisons between racial groups.

Age misreporting occurs in both death reports and census data. Because deaths and population are distributed differently by age, age errors cause different patterns of relative net error in death counts by age (numerators in estimated rates) and in population counts by age (denominators). The volume of deaths and risk population that are reallocated to incorrect ages depends on age structure as well as on rates of age misreporting. This led Bhat (1990) to argue that models should include gross, rather than net, rates of age misstatement. The question of how deaths and exposure may be reported at incorrect ages is the focus of Preston, Elo, and Stewart (1999), who consider the differential effects of age under- and over statement on mortality and life expectancy. (Preston, Elo, and Stewart 1999).

Several recent studies address another main data quality problem: under-registration of deaths and under-enumeration of risk populations. Many authors have developed methods to produce estimates of infant and child mortality from limited or defective data (Hill 1991; Hill, Choi, and Timæus 2005; Romero Prieto, Verhulst, and Guillot 2021). Palloni and Pinto-Aguirre (2011) and Beltrán-Sánchez et al. (2020) find significant errors in age reporting and registration for a series of countries in Latin America. A large body of literature has documented incomplete registration in low- and middle-income countries, and in subnational regions (Peralta et al. 2019; Castanheira and Monteiro da Silva 2022; Gupta and Mani 2022; Ouedraogo 2020). Schmertmann and Gonzaga (2018) propose a probabilistic method for correcting under-registration of deaths. Glei, Barbieri, and Santamaría-Ulloa (2019) examine the quality of mortality data in Costa Rica using a variety of demographic methods to identify errors related to digit preference, age overstatement, and completeness of death registration. They find that old-age mortality estimates in Costa Rica in the 1970s and 1980s were biased downward due to incomplete registration and to age declaration errors. Glei et al. (2021) evaluate mortality data quality in Mexico since 1990 and find clear signs of age heaping on death reports before 2000. Li and Gerland (2013) propose an indirect method to estimate old-age mortality based on census data.

In sum, there is a rich empirical literature on data errors in mortality estimates, but demographers know less about the formal analytics of how multiple sources of error combine and interact to cause bias. We focus on that gap in this paper.

3. Notation and estimation bias

3.1 Notation

Suppose that there are A age groups that start at integer ages $y = 0, 1, \ldots (A-1)$, where the last interval may be open. Call c_y the probability that a living y-year-old appears in official population counts (whether or not their age is reported correctly), and denote p_{xy} as the probability that a counted individual with a true age y reports their age as x. Define an $A \times 1$ vector of age-specific census coverage probabilities $c = (c_0 \ldots c_{A-1})'$ and an $A \times A$ age reporting matrix P with p_{xy} in the xth row and yth column. Define an analagous vector v for age specific registration of deaths and matrix Q for age misreporting on death certificates.

The $A \times 1$ vectors of reported population and death counts by age (denoted n and d, respectively) are related to the vectors of true counts (η and δ) by

$$\begin{aligned} n &= P\,\text{diag}\,(c)\;\eta \\ d &= Q\,\text{diag}\,(v)\;\delta\;. \end{aligned} \tag{1}$$

For the population the $A \times 1$ vector of estimated mortality rates by (possibly misreported) age will be

$$m = (\tfrac{d_0}{n_0}\;\cdots\;\tfrac{d_{A-1}}{n_{A-1}})' = [\text{diag}\,(n)]^{-1}\,d\,. \tag{2}$$

Data errors occur when $\text{diag}\,(c) \neq I$ (imperfect census coverage), $P \neq I$ (imperfect census age misreporting), $\text{diag}\,(v) \neq I$ (imperfect death registration in vital statistics), $Q \neq I$ (imperfect age reporting on death certificates), or any combination of these.

3.2 Bias in estimated mortality rates

We first want to understand bias in mortality rate estimates – that is, how the vector m changes when $\text{diag}\,(c)\,, P, \text{diag}\,(v)$, and/or Q are not all equal to identity matrices I. Call ϵ a generic scalar parameter that affects one or all of the multiplier matrices. From Equation (2) the derivative of the vector of estimated mortality rates with respect to ϵ is

$$\begin{aligned} m_\epsilon &= \tfrac{\partial}{\partial\epsilon}\left([\text{diag}\,(n)]^{-1}\right) d + [\text{diag}\,(n)]^{-1}\,d_\epsilon \\ &= [\text{diag}\,(n)]^{-1}\,[-\text{diag}\,(n_\epsilon)]\,[\text{diag}\,(n)]^{-1} d + [\text{diag}\,(n)]^{-1} d_\epsilon \\ &= \text{diag}\left(\tfrac{1}{n}\right)\{d_\epsilon - [\text{diag}\,(n_\epsilon)]\;m\}\;, \end{aligned} \tag{3}$$

where the ϵ subscript represents the derivative of a vector or matrix with respect to ϵ.

The changes in n and d in Equation (3) are

$$\begin{aligned} n_\epsilon &= [P_\epsilon\,\text{diag}\,(c) + P\,\text{diag}\,(c_\epsilon)]\,\eta \\ d_\epsilon &= [Q_\epsilon\,\text{diag}\,(v) + Q\,\text{diag}\,(v_\epsilon)]\,\delta\,. \end{aligned} \tag{4}$$

In the calculations that follow we analyze the effects of data errors when we start at perfect reporting ($P = Q = \text{diag}\,(c) = \text{diag}\,(v) = I$) and introduce small errors. By using the derivative formulas, starting from perfect reporting, we are implicitly considering the consequences of very small inaccuracies. We can also estimate the effects of simultaneous small changes in elements of P, $\text{diag}\,(c)$, Q, and $\text{diag}\,(v)$ by adding their derivatives.

Combining and rearranging Equation (3) and Equation (4) in the case of perfect reporting, and using μ to designate true mortality rates, we can decompose changes in the vector of age-specific mortality rates as

$$
\begin{aligned}
m_\epsilon\big|_{\text{perf. rep.}} = \; & + \text{diag}\left(\tfrac{1}{\eta}\right)\, Q_\epsilon\, \delta && \text{(death age misreporting)} \\
& + \text{diag}\left(\mu\right)\, v_\epsilon && \text{(death coverage)} \\
& - \text{diag}\left(\tfrac{\mu}{\eta}\right)\, P_\epsilon\, \eta && \text{(census age misreporting)} \\
& - \text{diag}\left(\mu\right)\, c_\epsilon\,, && \text{(census coverage)}
\end{aligned}
\tag{5}
$$

where $\text{diag}\left(\frac{\mu}{\eta}\right) = \text{diag}\left(\frac{\mu_0}{\eta_0} \cdots \frac{\mu_{A-1}}{\eta_{A-1}}\right)$.

3.3 Bias in estimated life expectancy

In Appendix Equation (33) we show that the derivative of estimated life expectancy with respect to a change in the mortality rate at single-year age y is well approximated by

$$
\frac{\partial e_0}{\partial m_y} = -\bar{T}_y\,, \tag{6}
$$

where $\bar{T}_y = \frac{1}{2}(T_y + T_{y+1})$. Thus the total effect of reporting errors on e_0 would be

$$
\frac{\partial e_0}{\partial \epsilon} = -\bar{T}'\, m_\epsilon\,, \tag{7}
$$

with $\bar{T}' = (\bar{T}_0\ \bar{T}_1\ \bar{T}_2\ \ldots)$.

4. Types of reporting errors

We begin by identifying three simple types of errors: under-registration of deaths, under-enumeration of population, and misreports of y-year-olds as x-year-olds. We analyze the effects of each error type separately before considering their combined effects.

4.1 Undercounts of deaths and/or population

We first consider the effects of under-registration of deaths and/or under-enumeration of population at a single true age y. Suppose that the probabilities of a y-year-old being counted in the census, or of a death at true age y being registered, change as

$$\begin{aligned} c_y &\rightarrow (c_y - k_C \cdot \epsilon) \\ v_y &\rightarrow (v_y - k_V \cdot \epsilon) \,, \end{aligned} \tag{8}$$

where k_C and k_V may be different depending on the degree of under-registration in census versus vital registration. In matrix terms this change is

$$\begin{aligned} P_\epsilon &= 0 \qquad & c_\epsilon &= -k_C \, e_y \\ Q_\epsilon &= 0 \qquad & v_\epsilon &= -k_V \, e_y \,, \end{aligned} \tag{9}$$

where e_y is the yth column of the $A \times A$ identity matrix I. After a few algebraic steps (not shown) the bias in mortality rates from Equation (5) is

$$m_\epsilon \big|_{\text{perf. rep.}} = (k_C - k_V) \, \mu_y \, e_y \,. \tag{10}$$

This means that under-registration at age y affects mortality estimates at age y only, with

$$\frac{\partial m_y}{\partial \epsilon} \big|_{\text{perf. rep.}} = (k_C - k_V) \, \mu_y \,, \tag{11}$$

and mortality unchanged at all other ages. If omissions are greater in the census ($k_C > k_V$), then estimated mortality at age y is biased upwards. If omissions are greater in vital statistics ($k_V > k_C$), then the bias is downward. If omissions are equal ($k_C = k_V$), then under-registration does not cause bias.

Given perfect initial reporting, the impact of age-specific under-registration on life expectancy is also simple:[4]

$$\frac{\partial e_0}{\partial \epsilon}\Big|_{\text{perf. rep.}} = (k_V - k_C)\,\bar{T}_y\,\mu_y\,. \tag{12}$$

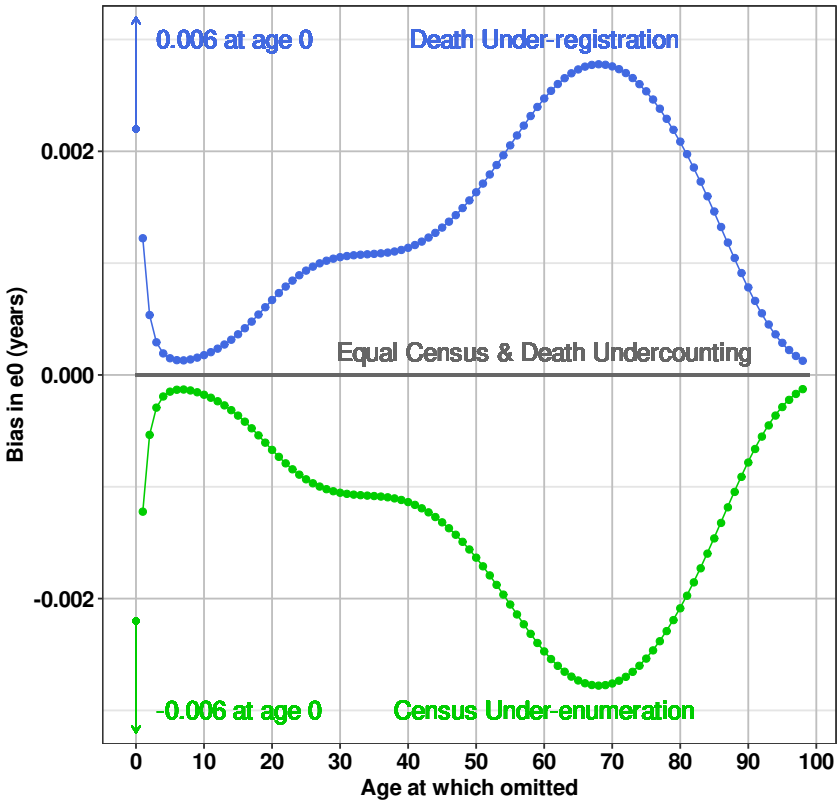
Figure 1 provides an example of empirical calculations with under-registration, using period data from males in the Brazilian state of São Paulo in 2009–2011.[5] In the figure we treat the original death and exposure data as complete and calculate the bias in e_0 caused by a 1% increase in age-specific under-registration in census counts (lower line), in death counts (upper line), or in both (flat center line) in a stationary population with São Paulo male death rates by age.

The calculations in Figure 1 show that the effects of under-registration or under-enumeration vary considerably over ages. In general, ages at which deaths are concentrated tend to be those at which omissions will cause that greatest bias in e_0. Because of the interaction between the density of deaths and remaining life expectancy, however, the age of peak omission bias (apart from infancy) will be lower than the modal age at death.

[4] An interesting corollary of Equation (12) is that the effect of either type of under-registration on life expectancy will be greatest near ages y, at which there are approximately 10 years of remaining life ($e_y \approx 10$). In Section A-2 we show that $T(a)\mu(a)$ is an increasing function of age at adult ages a, where remaining life expectancy satisfies $e(a) > \frac{1}{b}$ and b is the age derivative of the log mortality rate. Because $b \approx 0.10$ in most human populations, the bias in life expectancy estimates caused by under-registration will therefore reach an extreme (a minimum if census under-registration is greater than death under-registration, or a maximum if death under-registration is greater) at the age where remaining life expectancy is approximately 10 years. For most modern human populations this age of maximum bias would fall between 70 and 80.

[5] In order to de-emphasize coincidental population features and focus on more universal patterns, we have smoothed the schedule of log mortality rates over ages 0 to 99 and extrapolated to ages 100 to 119 using the Kannisto approach (Thatcher, Kannisto, and Vaupel 1998). All empirical calculations also use the stationary population associated with these age-specific mortality rates.

Figure 1: **Bias in e_0 from 1% under-registration of deaths or under-enumeration of population at different ages**



Note: Stationary population with São Paulo 2009–2011 male mortality rates.

4.2 Simple age misreporting in census and/or death registers

We now consider the effects of an increase in the proportion of registered deaths or enumerated population members at age y that are misreported as age x. For this analysis suppose that the probabilities in Equation (1) change as

$$
\begin{aligned}
p_{xy} &\to (p_{xy} + k_P \cdot \epsilon) \\
q_{xy} &\to (q_{xy} + k_Q \cdot \epsilon) \\
& \\
p_{yy} &\to (p_{yy} - k_P \cdot \epsilon) \\
q_{yy} &\to (q_{yy} - k_Q \cdot \epsilon)
\end{aligned}
\tag{13}
$$

so that there are (possibly different) increases in the fraction of y year-old-members of the recorded population who are misreported as x-year-olds and in the fraction of registered deaths to y-year-olds that are misreported as x, with compensating decreases in the probability of correct reporting for both.

In matrix notation the changes are

$$
\begin{aligned}
P_\epsilon &= k_P \cdot (e_x e_y' - e_y e_y') \qquad & c_\epsilon &= 0 \\
Q_\epsilon &= k_Q \cdot (e_x e_y' - e_y e_y') \qquad & v_\epsilon &= 0\,.
\end{aligned}
\tag{14}
$$

Substituting these changes into Equation (5) yields (again with some omitted steps)

$$
m_\epsilon\big|_{\text{perf. rep.}} = \begin{bmatrix} \vdots \\ (k_P - k_Q)\,\mu_y \\ \vdots \\ \left(\frac{\eta_y}{\eta_x}\right)(-k_P\,\mu_x + k_Q\,\mu_y) \\ \vdots \end{bmatrix} \quad , \tag{15}
$$

with the first term in the yth position and the second in the xth position, and all other vector elements equal to zero.

In other words, $y \to x$ misreporting affects estimated mortality at ages x and y only. Census misreporting $y \to x$ causes positive bias in estimated mortality at age y and negative bias at age x. Misreporting on death certificates $y \to x$ does the opposite.

The bias in estimated life expectancy at birth caused by $y \to x$ misreporting would therefore be

$$
\frac{\partial e_0}{\partial \epsilon}\Big|_{\text{perf. rep.}} = k_P \left[-\bar{T_y}\,\mu_y + \bar{T_x}\left(\frac{\eta_y}{\eta_x}\right)\mu_x\right] + k_Q \left[\bar{T_y}\,\mu_y - \bar{T_x}\left(\frac{\eta_y}{\eta_x}\right)\mu_y\right] . \tag{16}
$$

It is useful to consider the separate effects of census and death age misreporting from Equation (15) and Equation (16). Table 1 summarizes for the cases of census age misreporting only ($k_P = 1, k_Q = 0$), death age misreporting only ($k_P = 0, k_Q = 1$), and equal amounts of age misreporting in both sources ($k_P = 1, k_Q = 1$). Other combinations are possible, but these archetypes tell the main story.

Table 1: Effects of simple age misreporting on mortality and life expectancy

	Type of $y \to x$ age error		
	a. Census only $(k_P = 1, k_Q = 0)$	**b. Deaths only** $(k_P = 0, k_Q = 1)$	**c. Both** $(k_P = 1, k_Q = 1)$
$\Delta\mu_y$ (donor age)	$+\mu_y$	$-\mu_y$	0
$\Delta\mu_x$ (recipient age)	$-\left(\frac{\eta_y}{\eta_x}\right)\mu_x$	$+\left(\frac{\eta_y}{\eta_x}\right)\mu_y$	$+\left(\frac{\eta_y}{\eta_x}\right)(\mu_y - \mu_x)$
Δe_0	$-\bar{T}_y\mu_y + \left(\frac{\eta_y}{\eta_x}\right)\bar{T}_x\mu_x$	$+\bar{T}_y\,\mu_y - \left(\frac{\eta_y}{\eta_x}\right)\bar{T}_x\,\mu_y$	$+\left(\frac{\eta_y}{\eta_x}\right)\bar{T}_x(\mu_x - \mu_y)$
	Bias when $y < x$ (age overstatement) (if mortality rates increase with age)		
μ_y (donor age)	Positive	Negative	Zero
μ_x (recipient age)	Negative	Positive	Negative
e_0	Likely Positive*	Likely Positive*	Positive

Note: * Depends on population age structure.

A key point for understanding the life expectancy bias caused by age misreporting is that biases in mortality rates occur at multiple ages and partially offset one another. For example, $y \to x$ age overstatement on death certificates (column b in Table 1, with $y < x$) will cause positive bias in mortality over the (higher) recipient age interval $[x, x+1)$, but negative bias in mortality at the (lower) donor age interval $[y, y+1)$. The result in e_0 calculations is an upward bias in the probability of surviving to x but a downward bias in the expected number of years lived after x. The total effect on e_0 is not immediately obvious and requires some careful demographic thinking.

Analysis of the results in the Δe_0 row of Table 1 shows that age overstatement ($y < x$) would usually, but not always, bias life expectancy estimates upward. For age understatement ($x < y$) results are reversed. These bias patterns are for both census and death age overstatement and for combinations.

To understand patterns of bias in life expectancy better, consider the three cases a, b, and c in Table 1. The main demographic regularities to keep in mind are that for $y < x$

at adult ages we expect

$$\begin{aligned} \mu_y &< \mu_x && \text{(if mortality risks increase with age)} \\ T_y &> T_x \\ T_y\mu_y &< T_x\mu_x && \text{(unless } e_y < \approx 10\text{)} \\ \left(\frac{\eta_y}{\eta_x}\right) &> 1\,, && \text{(if population decreases with age)} \end{aligned} \tag{17}$$

where the next-to-last inequality comes from Appendix Section A-2. Notice that the last inequality, unlike the others, depends on the population's age structure.

Age overstatement in census data only (case a in Table 1 with $y < x$) would usually lead to positive bias in life expectancy. Over most adult ages, $T \cdot m$ increases with age and $\left(\frac{\eta_y}{\eta_x}\right)$ is usually greater than one, so the change in e_0 caused by census age overstatement is likely positive. There could be exceptions, however. Unusual age structure effects and a sufficiently low $\left(\frac{\eta_y}{\eta_x}\right)$ could reverse the direction of bias – if there were more exposure at the higher (incorrect, recipient) age x than at the lower (correct, donor) age y, then census age overstatement could cause negative bias in e_0.

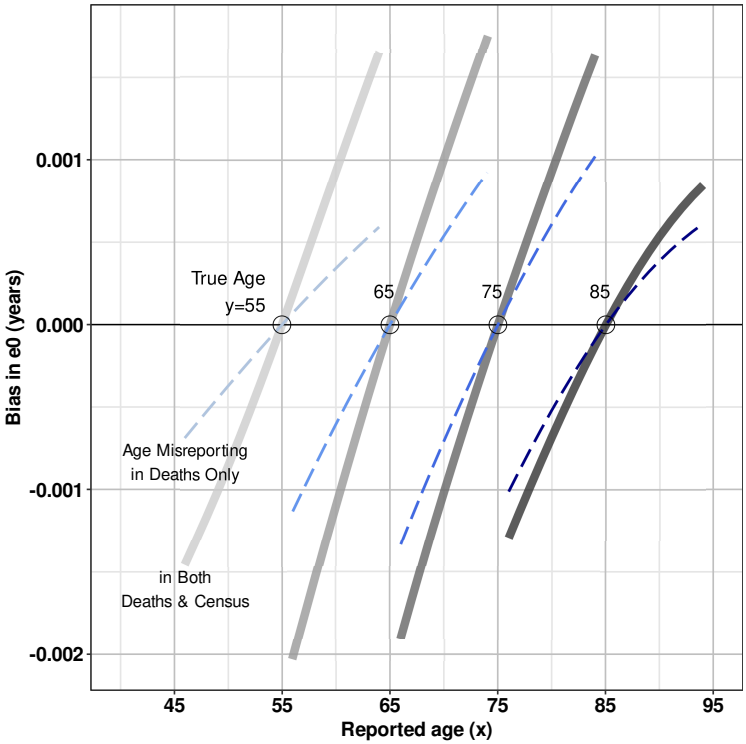
Age overstatement in death data only (case b in Table 1 with $y < x$) has a less certain effect on e_0. Because $T_y > T_x$ (the second inequality in Equation (17)) it will cause positive bias if the number of x-year-olds in the population is similar to or greater than the number of y-year-olds, $\left(\frac{\eta_y}{\eta_x}\right) < \left(\frac{T_y}{T_x}\right)$. However, if there are many more y-year-olds, $\left(\frac{\eta_y}{\eta_x}\right) > \left(\frac{T_y}{T_x}\right)$ then the bias in e_0 could be negative.

Equal age overstatement in both census and death data (case c in Table 1 with $y < x$) is certain to cause positive bias in life expectancy as long as mortality rates are higher at the higher, incorrect age x than at the lower, correct age y.

Figure 2 shows illustrative empirical calculations for the bias effects of misreporting of the ages of 55, 65, 75, and 85 on death certificates and in both death certificates and censuses equally. These calculations use a stationary population based on São Paulo male mortality rates and therefore do not contain any unusual age structure effects. As a result, in Figure 2 all biases in e_0 are positive for age overstatement and negative for understatement. The bias effects of by age misreporting in census counts, which are implicit in Figure 2 as the difference between the two curves for each true age y, also operate in the same direction: Overstatement of ages leads to overestimates of e_0, and vice-versa.[6]

[6] In the real São Paulo 2010 population, unlike the stationary version, there are several cases of counterintuitive bias. For example, death age overstatement $85 \rightarrow 93$ or $55 \rightarrow 64$ would cause small negative biases in e_0 for São Paulo males.

Figure 2: **Life expectancy bias if 1% of those with true age y have a reported age of x**



Notes: Vertical scale shows the error in estimated e_0 that occurs if all individuals are counted and all deaths are registered, but 1% of those with true age y are reported as age x. Dashed lines represent life expectancy bias when only death reports contain age errors. Solid lines represent combined effects when both census population and death reports have $y \rightarrow x$ errors. Stationary population with São Paulo 2009–2011 male mortality rates.

The primary conclusion from Figure 2 is that bias effects of age misreporting on life expectancy are quite small. Although the derivative calculations are valid only for very small levels of misreporting, they give a good sense of the order of magnitude of the bias that larger changes would cause. In practice, ages will be both under- and overstated (Bhat 1990; Palloni, Beltrán-Sánchez, and Pinto 2021; Preston, Elo, and Stewart 1999) so that even the total bias effect of much higher levels of misreporting at many different ages y would be measured in tenths of years, at most.

4.3 Age misreporting at many ages simultaneously

We now investigate more complex situations in which age misreporting occurs simultaneously at many ages, in both directions. In the examples that follow we use three patterns of single-year age misreporting derived from published demographic studies. These patterns are summarized in Table A-1. We emphasize that only the Palloni, Beltrán-Sánchez, and Pinto (2021) paper for the Costa Rican pattern actually includes single-year age misstatement probabilities. For the other two patterns in Table A-1 we fit a single-year age misstatement matrix to approximate published misstatement probabilities for five-year age groups, using a parameteric model similar to that of Palloni, Beltrán-Sánchez, and Pinto (2021). Appendix A-4 describes the model. The misreporting fractions reported for ages 60+ in Table A-1 are weighted using a stationary population with male 2010 mortality rates from São Paulo. From here on we use 'Costa Rican,' 'African American,' and 'Indian' as shorthand labels for these specific misreporting patterns, recognizing that they do not apply to all Costa Rican, African American, or Indian populations.

Table 2: Example patterns of age misstatement

Pattern/Abbrev	Costa Rica/CR	African American/AA	India/IN
Source	Palloni, Beltrán-Sánchez, and Pinto (2021)	Preston, Elo, and Stewart (1999)	Bhat (1990)
Type of data	Census both sexes	Female deaths	Census males
Reference period	2000s	1980s	1970s
Published misstatement probabilities	1-year	5-year group	5-year group
1-year rates from...	original paper	model fit by this paper's authors	model fit by this paper's authors
Description	overstatement more likely than understatement; small errors more likely than large	understatement much more likely than overstatement; large negative errors	almost all ages misstated; large errors likely in both directions
True Age 60+% understating: cond. mean error	15% : –2.8 years	24% : –5.5 years	40% : –4.4 years
True Age 60+% overstating : cond. mean error	25% : +1.9 years	3% : +1.3 years	60% : +4.2 years
Misstatement Matrix	Π_{CR}	Π_{AA}	Π_{IN}

4.3.1 Bias in mortality rates

We begin with a mathematical approach, by considering small changes from perfect reporting in the direction of one of the Π matrices in Table A-1. Because most available data is for age misreporting at higher adult ages only, we study bias in mortality rates above age 60 and e_x for higher ages rather than e_0.

For pattern $i \in \{CR, AA, IN\}$, define $\Delta_i = (\Pi_i - I)$, and suppose that a shift from perfect age reporting toward that pattern uses misreporting matrix $I + \epsilon \cdot \Delta_i$. For derivative calculations, use

$$\begin{aligned} P &= I + k_P \cdot \epsilon \cdot \Delta_i \\ Q &= I + k_Q \cdot \epsilon \cdot \Delta_i \,, \end{aligned} \tag{18}$$

with matrix derivatives

$$\begin{aligned} P_\epsilon &= k_P \cdot \Delta_i & c_\epsilon &= 0 \\ Q_\epsilon &= k_Q \cdot \Delta_i & v_\epsilon &= 0 \,, \end{aligned} \tag{19}$$

Substituting into Equation (5) yields

$$m_\epsilon\big|_{\text{perf. rep.}} = -k_P \,\text{diag}\left(\frac{\mu}{\eta}\right)(\Delta_i)\,\eta + k_Q \,\text{diag}\left(\frac{1}{\eta}\right)(\Delta_i)\,\delta \quad i \in \{CR, AA, IN\}, \tag{20}$$

from which we can approximate changes in log mortality rates as $[\text{diag}\,(m)]^{-1} m_\epsilon$.

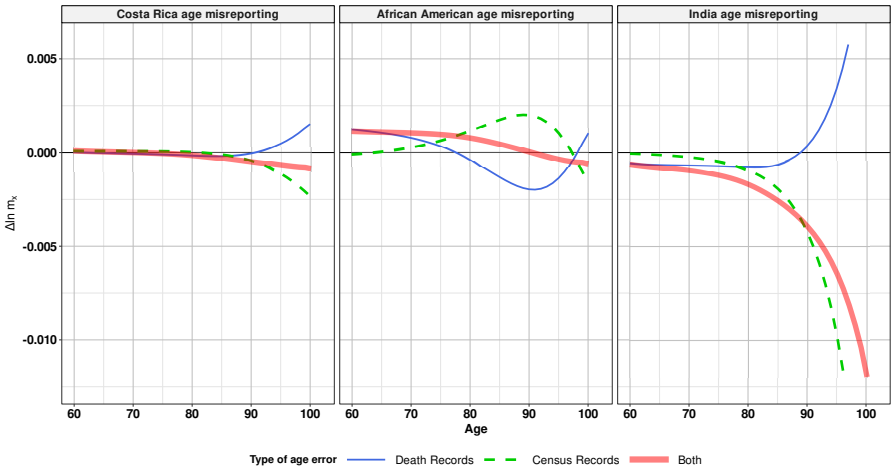
Figure 3 shows derivative calculations for the effects on log mortality rates of small errors in census age reporting, in death age reporting, or in both sources. If both death and census reports had misstated ages, then all three misreporting patterns would lead to negative biases in estimated mortality rates at ages above 90. However, there are interesting differences in age-specific and overall bias.

With the Costa Rican pattern of age misstatement, age errors on death certificates would tend to raise estimated death rates at ages 90+. The same errors on census reports would tend to lower estimated mortality at ages 80+. Combined age misreporting on both death and census records would cause almost no bias in estimated rates at ages below 80 and small negative biases at ages 80+.

With the African American pattern, in which underreported ages are much more common, misreporting on census records would cause positive bias in estimated mortality at all but the highest ages. In constrast, age misstatement on death records would cause positive bias at ages below 80 and above 95, but negative bias between those ages. The combined result of these errors would be overestimates of mortality rates at ages below 90 and small underestimates at ages 90+.

With the Indian pattern, which includes large misstatement errors in both directions, age errors on census records would cause a downward bias at all ages above 60, while age errors on death records would lead to a small downward bias at ages below 90 and a large upward bias at ages 90+. The combined effect of age errors on both sources would be a large downward bias in mortality rate estimates at all advanced ages.

Figure 3: Mortality rate bias caused by age misreporting on 1% of death or census records, using derivative formula



Notes: Derivative formula Equation (20) with $(k_P = 0, k_Q = .01)$ for death age misreporting, $(k_P = .01, k_Q = 0)$ for Census age misreporting, and $(k_P = .01, k_Q = .01)$ for misreporting in both sources. Alternative age misreporting patterns described in Table A-1. Stationary population with São Paulo 2009–2011 male mortality rates.

4.3.2 Net imports and exports of exposure and deaths by age

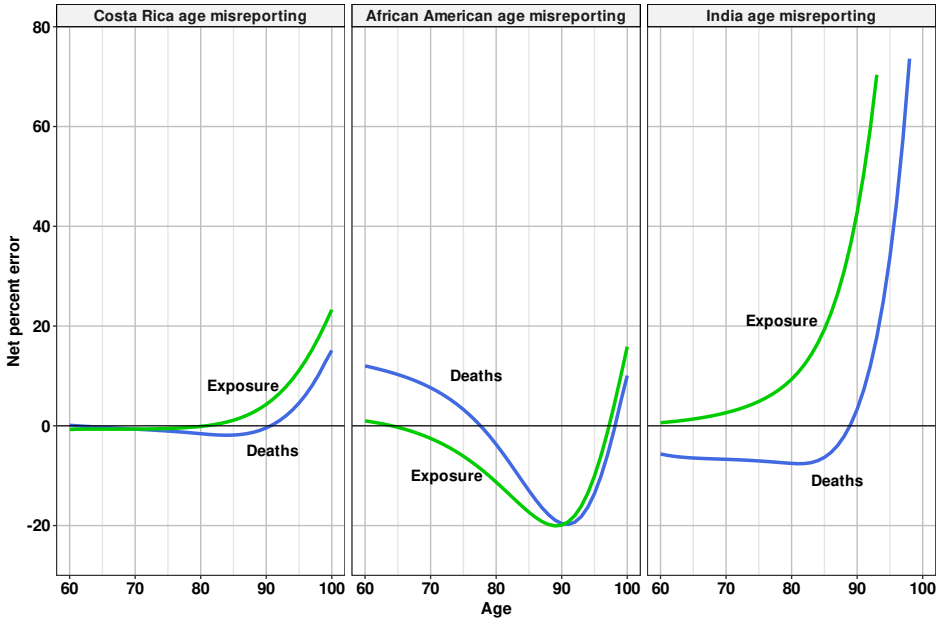
Derivative calculations such as Equation (20) are useful for predicting bias patterns, but it is important to understand their origins. For the case in which all deaths and person-years are counted, Figure 4 illustrates how incorrect age reports would affect total death and exposure counts by reported age for each of the three misreporting patterns.

As an example, consider the exposure at reported age 90 when the true population is our example stationary population. With the African American age misreporting pattern on census records, more true 90-year-olds would report that they are not 90 than vice versa. Thus age misreporting would lead to 'net exports' of exposure (in this case, a negative net error of approximately 20%) at age 90. With the Costa Rican or Indian patterns this result is reversed: The reported population at age 90 would be higher than the true population.

Deaths are reallocated between true and false ages in the same way, leading to the biases illustrated in Figure 4. For example, with the African American pattern more deaths to 90-year-olds are reported at other ages than vice versa, resulting in an underestimate of d_{90} of about 20%. Because African American age errors would lead to equiproportional errors in both deaths and exposure at age 90, the estimated mortality rate would have

almost no bias despite the age errors (cf. the derivative calculation in Figure 3, middle panel at age 90).

Figure 4: Sources of bias in mortality rate estimates



Notes: Net percent errors in death and exposure, by reported age, when there is complete death registration and complete census enumeration, but all death and census reports are subject age misreporting. Alternative age misreporting patterns described in Table A-1. Stationary population with São Paulo 2009–2011 male mortality rates.

For mortality rates, the end result of these imports and exports of deaths and exposure across ages are the age-specific biases illustrated in Figure 3. With Costa Rican misreporting there are only small net errors in age-specific exposure and deaths at ages below 80. Above 80 net overcounts of exposure are only slightly larger than overcounts of deaths. As a result, age misstatement of this type would cause small biases in estimated mortality rates, mostly at ages above 80. With African American age misreporting, in contrast, a large fraction of elders will understate their true age. At ages below about 78 this causes underestimated exposure, overestimated deaths, and overestimates of mortality rates. At higher ages the interactions between net errors in exposure and deaths are more complicated, resulting in decreasing positive bias between ages 80 and 90, and a

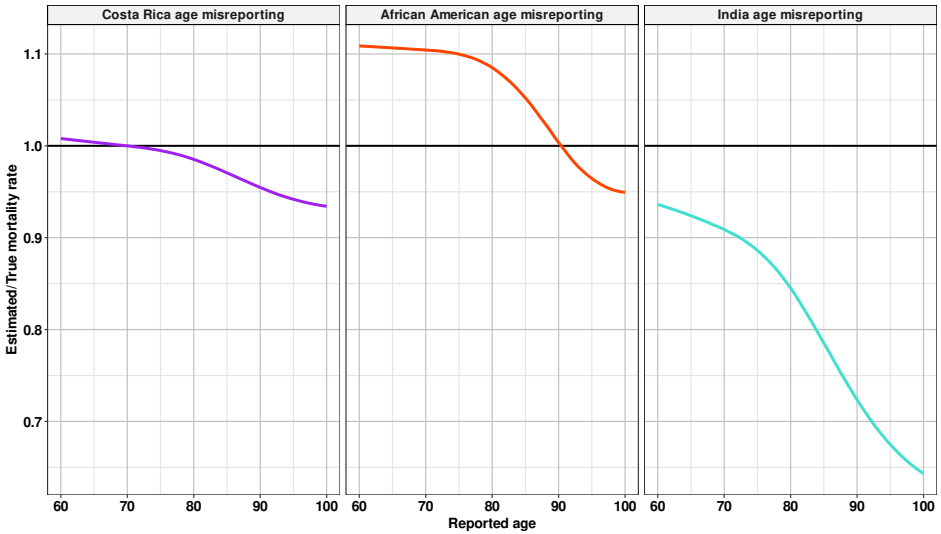
small negative bias at ages 90+.[7] Last, with the Indian pattern of large age reporting errors in both directions there would be net positive errors in exposure at all ages 60+, net negative errors in deaths below the high 80s, and net positive errors at higher ages. When combined these errors would cause a negative bias in estimated mortality rates at all ages 60+, with especially large underestimates at the highest ages.

4.3.3 Ratio of estimated/true mortality

Extrapolation of the small changes in Figure 3 could produce large approximation errors if there are important nonlinearities or interactions between parameters. In order to address this concern, Figure 5 supplements the analytical derivatives with calculations that use the full Π matrices for the three misreporting patterns under the assumption that all census and death records, rather than only 1%, are subject to age misreporting. These calculations confirm the utility of the derivative formula: For each misreporting pattern, age-specific bias in mortality estimates matches well with what one would expect from Equation (20) and Figure 3. With Costa Rican misreporting the ratio of estimated/true mortality is close to one below age 70, and falls to approximately 0.93 over ages 70 to 100. With African American misreporting mortality rates will be overestimated at ages below 90, and the ratio of estimated/true mortality falls to 0.95 by age 100. With the much more pervasive age misreporting in the Indian pattern, estimated mortality would be approximately 94% of the true rate at age 60, falling steadily to approximately 64% at age 100.

[7] Net imports of deaths and exposure at the highest ages, with positive bias in both d_x and n_x (but larger proportional positive bias in n_x), is described in Preston, Elo, and Stewart (1999). In our example stationary population that argument applies only to the very highest ages – approximately 98+.

Figure 5: **Ratios of estimated to true mortality, by age, if all census and death records are subject to age misstatement**
($P = Q = \Pi_i,\ i \in \{CR, AA, IN\}$)



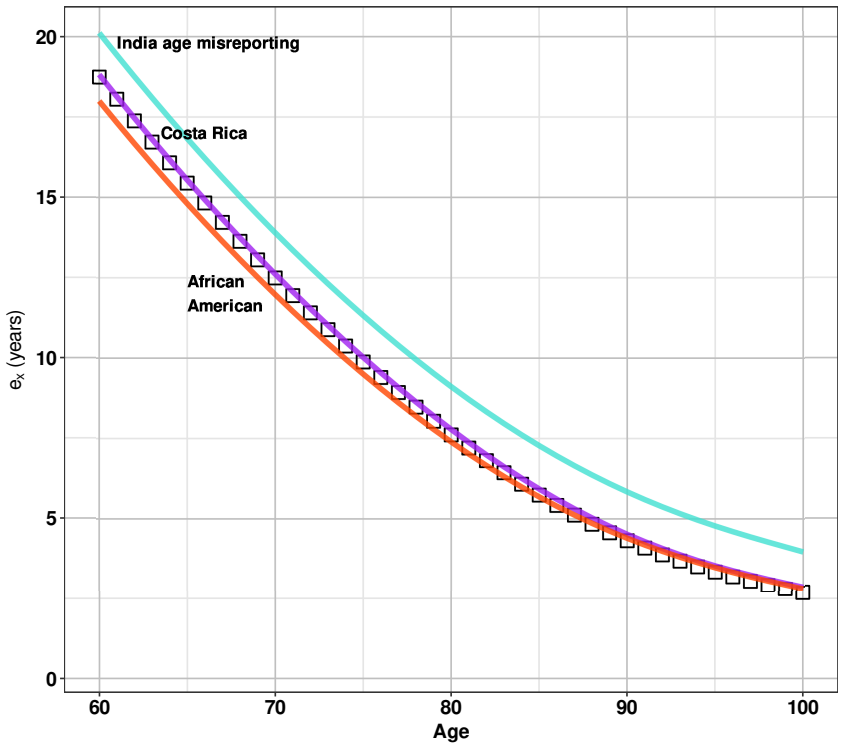
Notes: Alternative age misreporting patterns described in Table A-1. Stationary population with São Paulo 2009–2011 male mortality rates.

4.3.4 Bias in e_x at ages above 60

Because we often measure and compare population mortality levels using remaining life expectancies, it is important to understand whether the biases in mortality rates caused by age misstatement would significantly change summary indices such as e_{60} or e_{80}.

Figure 6 addresses this question by calculating remaining life expectancy e_x at all ages 60+ in our example stationary population. Squares represent true e_x values for a population with the assumed mortality rates. These decline from 18.7 years at age 60 to 2.7 years at age 100.

Figure 6: Estimated remaining life expectancy at ages 60+ with age misreporting on both census and death records



Notes: Calculations assume complete coverage and complete enumeration. Squares represent true e_x values; lines correspond to estimates with different age misstatement patterns. Alternative age misreporting patterns described in Table A-1. Stationary population with São Paulo 2009–2011 male mortality rates.

From these calculations we conclude that age misstatement likely has only small effects on estimates of remaining life expectancy. With the Costa Rican pattern of small, nearly symmetric errors and low m_x bias at advanced ages, e_x bias is very small indeed – all estimates are within 0.2 years of the correct values. With the African American misreporting pattern mortality rates are overestimated at ages 60 to 89, and consequently remaining life expectancies are underestimated. Even so, the largest absolute error in estimated e_x with African American misreporting is 0.7 years, at age 60.[8] Biases in e_x are notably larger, and always positive, with the Indian pattern of age misstatement. How-

[8] True e_{60} in our example is 18.7 years. With African American misreporting the estimated value would be 18.0.

ever, even with the large downward biases in estimated mortality rates seen in Figures 3 and 5, the upward bias in e_x is fairly modest, ranging from +1.3 to +1.5 years over this age range.[9]

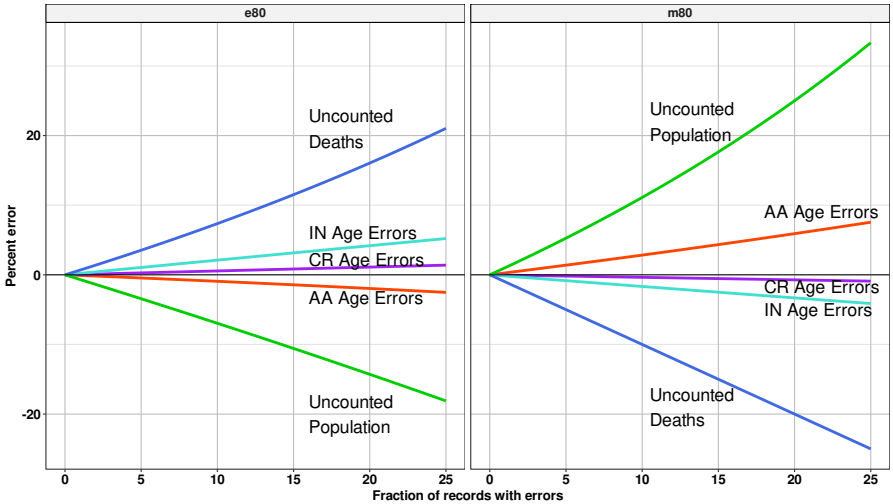
5. Comparative effects of undercounts and age misreporting

After analyzing each type of data error separately, it is useful to directly compare the directions and sizes of their effects. Figure 7 shows two examples using the stationary São Paulo male population. The left panel shows the bias caused in e_{80} by data errors of different types, and the right panel shows the corresponding biases in μ_{80}. In both panels the horizontal axis shows the fraction of records with a given error, and the vertical axis represents the relative bias. For example, at 25% on the right edge of both panels, vertical distances indicate the relative bias caused by a 25% undercount of registered deaths at all ages (blue line), the relative bias caused by a 25% under-enumeration of exposure at all ages (green line), or when 25% of ages in both the census and death records are misreported (one line for each misreporting pattern in Table A-1).

Figure 7 illustrates one of our main findings: For life expectancy and mortality calculations, registration-related errors have far more influence than age misreporting errors. Omitting data has a much bigger effect than reshuffling the ages at which we record it.

[9] We consider these to be 'modest' errors because age misreporting is often blamed for much larger discrepancies in old-age life expectancy. Here we see that even big, nearly universal age reporting errors may not generate very large errors in e_x.

Figure 7: Relative bias in e_{80} and μ_{80} caused by different fractions of records with errors



Notes: Horizontal axis represents the fraction of deaths (at all ages) that are unregistered, the fraction of the population (at all ages) that is not enumerated, or the fraction of all deaths and exposure with misreported ages. Alternative age misreporting patterns described in Table A-1. Stationary population with São Paulo 2009–2011 male mortality rates.

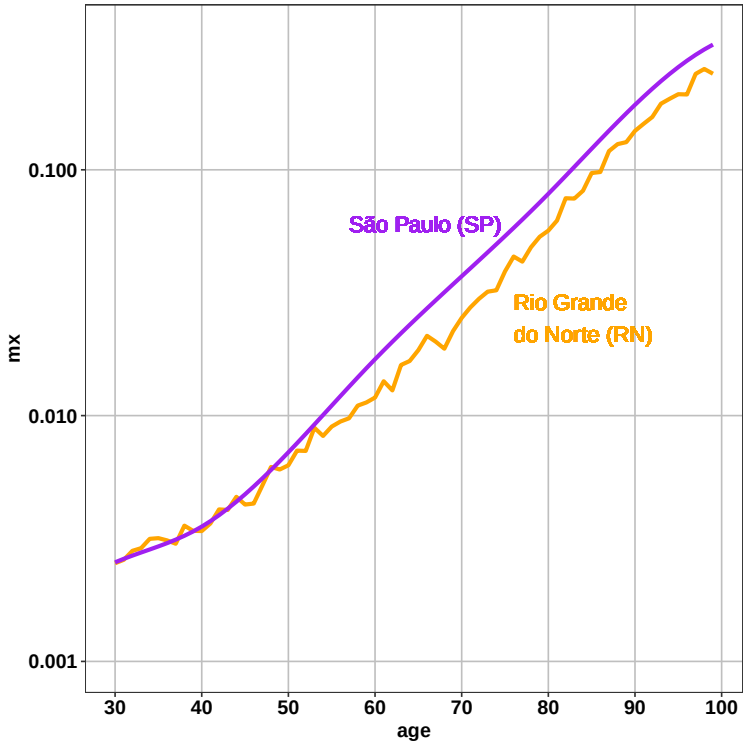
6. Misreporting and mortality crossovers

6.1 Example crossover

We can use the analytical framework developed here to consider the possible effects of different age-reporting errors and under-registration on comparative mortality patterns. In particular, we can consider the magnitude of errors that would be required to generate an observed crossover in age-specific mortality rates between two populations with identical mortality rates.

Figure 8 provides an example crossover, showing estimated mortality rates for males in the Brazilian states of São Paulo (SP) and Rio Grande do Norte (RN) in 2009–2011. SP male rates are lower until about age 40, after which the pattern reverses and RN mortality is lower. This crossover is suspicious, however, because RN is a much poorer, more rural states and we might reasonably expect it to have higher mortality rates at all adult ages.

Figure 8: São Paulo (SP) and Rio Grande do Norte (RN) male mortality rates by age, 2009–2011



Note: Logarithmic vertical scale.

One classic question about a crossover like this is whether it could be caused by data errors in the population with the lower rates at advanced ages. Here we illustrate how we can use our analytical framework to investigate this question. In our illustration we consider how errors in (P, Q, v, c) for one population (in our case, RN) might generate the observed crossover, even if true mortality rates were those of the other population.

6.2 Analysis: Data errors and false crossovers

Inverting the relationships in Equation (1) to write true exposure and events as a function of observed quantities produces

$$\begin{aligned} \eta &= [\text{diag}\,(c)]^{-1} P^{-1} n \\ \delta &= [\text{diag}\,(v)]^{-1} Q^{-1} d\,. \end{aligned} \tag{21}$$

Denoting e'_x as a $1 \times A$ row vector with a 1 in the position corresponding to age x and 0s elsewhere, the true exposure and deaths at age x are

$$\begin{aligned} \eta_x &= \tfrac{1}{c_x}\, e'_x P^{-1} n \\ \delta_x &= \tfrac{1}{v_x}\, e'_x Q^{-1} d\,. \end{aligned} \tag{22}$$

Replacing δ_x with its expected value $\mu_x \eta_x$ and combining the quantities in Equation (22) yields[10]

$$v_x = \left(\frac{e'_x Q^{-1} d}{\mu_x\, e'_x P^{-1} n} \right) c_x \qquad x = 0, 1, \ldots (A-1)\,. \tag{23}$$

Finally, stacking over ages produces

$$v = R\, c\,, \tag{24}$$

where R is an $A \times A$ diagonal matrix with the right-hand multipliers in Equation (23) as its diagonal elements.

Equation (23) and Equation (24) are alternative ways of expressing the expected relationship between reported d and n, true mortality rates μ, and reporting errors (P, Q, c, v). We can use these relationships to investigate crossovers. In particular, a false crossover could occur if true mortality rates were μ, estimated rates were $m = [\text{diag}\,(n)]^{-1} d$, and errors (P, Q, c, v) satisfied this mathematical relationship.

In the examples below we assume specific values for P and Q based on our different misreporting patterns, and then use Equation (24) in two specific cases: (1) to find the vector of age specific death coverage v required to generate the observed crossover when census coverage is perfect ($c = \iota$), and (2) to find, among coverage vectors c and v that generate the crossover, those that are closest to perfect coverage $\iota = (1 \cdots 1)'$.

[10] If $P = Q = I$ then Equation (23) simplifies to $\frac{v_x \mu_x}{c_x} = \frac{d_x}{n_x}$. This illustrates that observed (d_x/n_x) ratios do not identify mortality rates μ_x, even if reported ages x are accurate. Instead they identify $\mu_x \cdot (v_x/c_x)$ ratios.

Given age misreporting patterns P and Q, any pair of coverage vectors (c, v) that satisfy Equation (24) will generate estimates that match the crossover. Under our first criterion, the coverage levels necessary to generate the crossover if census reporting is complete ($c = \iota$) are

$$\begin{aligned} \tilde{c} &= \iota \\ \tilde{v} &= R\,\iota\,. \end{aligned} \tag{25}$$

Under the second criteron, the pair of coverage vectors that is closest to (ι, ι), in the sense of minimizing squared differences $(c-\iota)'(c-\iota) + (v-\iota)'(v-\iota)$, is

$$\begin{aligned} c^* &= (I + R'R)^{-1}\,(I + R)\,\iota \\ v^* &= R\,c^*\,. \end{aligned} \tag{26}$$

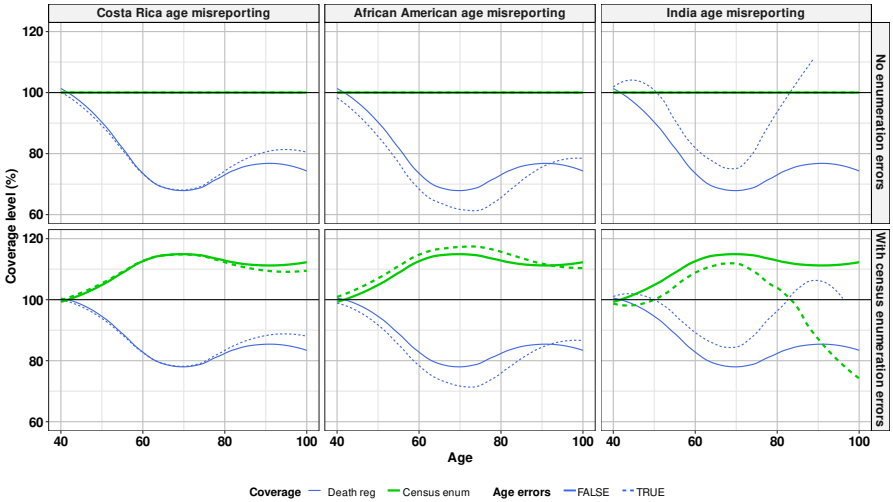
6.3 Example reporting errors that generate a crossover

The crossover in Figure 8 could be generated by under-registration of deaths, over-enumeration in the census, and/or age reporting errors in RN. We use Equation (24) to investigate how large those errors would have to be and how would they have to vary by age.

Even after we choose specific P and Q matrices for age misreporting, Equation (24) shows that there remain an infinite number of error patterns that would replicate the RN mortality rate estimates from the SP data. For illustrative purposes we focus on only a handful. In situations with imperfect census enumeration we use Equation (26) to find the c and v vectors that are closest to ι (i.e., perfect coverage) among all those that produce the observed crossover. For situations in which census enumeration is perfect ($c = \iota$) we use Equation (25) to find the v vector that would match the crossover. For each of these two coverage scenarios, we calculate first with perfect age reporting ($P = Q = I$, solid lines) and then with our three representative age misreporting patterns applied to both census and death registers ($P = Q = \Pi_i, i \in \{CR, AA, IN\}$, dashed lines).

Figure 9 shows error patterns in census enumeration and death registration by age that could generate the RN-SP crossover in Figure 8. In the top panels we assume that census coverage is 100% ($c = \iota$), in which case death registration coverage and misstated death and census ages would have to explain the crossover. In the bottom panels we also allow varying census coverage by age to affect a crossover. In all panels we calculate the coverage vectors (c, v) that satisfy Equation (24) when there are no age errors (solid lines), and when there is a specific pattern of age errors (Π_{CR}, Π_{AA}, or Π_{IN}) in both census and death records (dashed lines).

Figure 9: RN reporting errors that could generate the crossover in Figure 8, if RN rates were in fact identical to SP



Notes: Top panels show death registration levels needed to generate the crossover, if census coverage is perfect. Bottom panels show the crossover-generating coverage levels that are as close as possible to 1, if we allow counting errors in both census and deaths. Solid lines for cases with perfect age reporting are identical across columns; dashed lines for cases with age reporting errors differ across columns depending on the age misstatement pattern. Alternative age misreporting patterns described in Table A-1.

The solid lines in the top panels of Figure 9 show that if census enumeration is accurate and ages are correctly reported, then death registration coverage at ages 60+ in RN would have to be approximately 75% in order to generate a false crossover. The solid lines in the bottom panels show that if the census had RN overcounted residents 60+ by 10% to 15% while ages were accurately reported, then a false crossover could arise if approximately 80% of RN deaths at 60+ were registered.

The levels of death under-registration necessary to generate a crossover are highly implausible. Recent research in Brazil (Adair and Lopez 2018; Queiroz et al. 2020) includes estimated registration levels above 90% in RN in 2010, and there has been evidence of further improvements since then (Costa et al. 2020; Gonzaga et al. 2022).

Census over-enumeration in RN could contribute to a crossover, but it is even less likely. In a post-2000 census survey IBGE (2003) estimates enumeration errors for Brazilian states that ranged from –1% to –8%, and the estimated error for those 60+ in RN was –2.5%. Recent official projections for Brazilian states IBGE (2018) assume that 2010 census coverage errors were close to –1%.

If we introduce age misstatement as a possible additional source of a false crossover, the picture changes only a little. With the Costa Rican misstatement pattern a false crossover could arise at slightly higher levels of death registration coverage in RN (mainly at ages 80+), but the required levels are still well below the recent estimates. With the African American pattern of age misstatement (in which ages are frequently understated) a false crossover would require even lower and less plausible levels of RN death registration coverage at ages 60 to 90. The situation is more complicated if we assume the Indian age misstatement pattern: Although there could be a false crossover at higher levels of death registration coverage, it would require overregistration of deaths at the highest ages, which also seems highly implausible.

In sum, our analytical framework helps to show that the levels of counting and age reporting errors in RN that would be necessary to generate the crossover in Figure 8 are very implausible. Data errors might contribute to the crossover, but they are not a complete explanation.

7. Conclusion

We have constructed a mathematical model that incorporates coverage levels and age misreporting in both risk populations and deaths. Using this model as an analytical framework helps us to understand not only the effects of different types of reporting errors but also their relative magnitudes.

Previous literature focused mainly on empirical analysis and single cases from specific countries. We have added to this literature by using a mathematical approach to understand the net impact of data errors on mortality and life expectancy estimates.

Among our findings, we highlight that

- counting errors (death under-registration or census under-enumeration) are likely to cause much greater biases than age reporting errors,
- in the absence of counting errors, age misreporting in death and census data is unlikely to cause large biases in life expectancy estimates, and
- life expectancy estimates are most sensitive to under-registration of exposure or deaths at ages just before the modal age of life table deaths.

The patterns of bias caused by age misreporting evident in our mathematical calculations are broadly consistent with the empirical results in the pioneering work of Coale and Kisker (1990); Preston, Elo, and Stewart (1999), and others. Standard estimators tend to underestimate mortality rates even when age errors are symmetric (equal probability of under- or overstatement), bias increases with reported age, and biases become truly important only at the oldest ages.

An explicit mathematical framework allows us to calculate the kinds of reporting errors necessary to generate a 'false crossover.' This has been an important question in the demographic literature (Coale and Kisker 1986; Preston et al. 1996; Dowd and Hamoudi 2014), and it is valuable to have analytical tools to address the issue. In our Brazilian example, mathematical analysis does not explain the origin of the crossover, but it does allow us to virtually eliminate bad data for the less advantaged population as the only cause. Plausible age-reporting errors, death under-registration, or census enumeration errors would not be sufficient to generate the observed differences in mortality rates. The source of the crossover must be something else, such as very strong selection effects (Coale and Kisker 1986), real mortality differences, or both.

Our analysis also illustrates the importance of analyzing age-specific differentials in under-registration of deaths and under-enumeration of the risk population. At some ages these problems have virtually no impact on demographic calculations, while at others they can cause large biases.

The combination of mathematical and empirical approaches leads to several practical implications for demographers working in countries with defective registration and enumeration data. These mirror our analytical findings: Biases caused by data errors vary considerably across ages; age misreporting has fairly small effects on life expectancy calculations at old ages; and most importantly, enumeration and registration errors are likely to cause much bigger problems than age misreporting.

Our results reinforce the importance of continuous investment in civil registration and vital statistics (CRVS) systems in middle- and low-income countries. Advances in demographic methods can improve mortality estimates, but they are not substitutes for good quality CRVS systems.

8. Acknowledgments

Carl P. Schmertmann thanks the US Fulbright Scholar Program for financial support during the paper's conception and completion. Bernardo L. Queiroz and Marcos R. Gonzaga gratefully acknowledge support from the Brazilian National Research and Development Council (CNPq, fellowships 303928/2022-0 and 309661/2021-8, respectively).

References

Adair, T. and Lopez, A.D. (2018). Estimating the completeness of death registration: An empirical method. *PloS One* 13(5): e0197047. doi:10.1371/journal.pone.0197047.

Beltrán-Sánchez, H., Palloni, A., Pinto, G., and Verhulst, A. (2020). The Latin American Mortality Database (LAMBdA): Methodological document, version II, January 2020. [electronic resource]. www.ssc.wisc.edu/cdha/latinmortality2.

Bennett, N.G. and Horiuchi, S. (1981). Estimating the completeness of death registration in a closed population. *Population Index* 47(2): 207–221. doi:10.2307/2736447.

Bhat, P.M. (1990). Estimating transition probabilities of age misstatement. *Demography* 27(1): 149–163. doi:10.2307/2061559.

Camarda, C.G., Eilers, P.H., and Gampe, J. (2008). Modelling general patterns of digit preference. *Statistical Modelling* 8(4): 385–401. doi:10.1177/1471082X0800800404.

Castanheira, H.C. and Monteiro da Silva, J.H.C. (2022). Examining sex differences in the completeness of Peruvian CRVS data and adult mortality estimates. *Genus* 78(3). doi:10.1186/s41118-021-00151-5.

Coale, A.J. and Kisker, E.E. (1986). Mortality crossovers: Reality or bad data? *Population Studies* 40(3): 389–401. doi:10.1080/0032472031000142316.

Coale, A.J. and Kisker, E.E. (1990). Defects in data on old-age mortality in the United States: New procedures for calculating mortality schedules and life tables at the highest ages. *Asian and Pacific Population Forum* 4(1): 1–31.

Coale, A.J. and Li, S. (1991). The effect of age misreporting in China on the calculation of mortality rates at very high ages. *Demography* 28(2): 293–301. doi:10.2307/2061281.

Condran, G.A., Himes, C.L., and Preston, S.H. (1991). Old-age mortality patterns in low-mortality countries: An evaluation of population and death data at advanced ages, 1950 to the present. *Population Bulletin of the United Nations* 30: 23–60.

Costa, L.F.L., de Mesquita Silva Montenegro, M., Rabello Neto, D.d.L., de Oliveira, A.T.R., Trindade, J.E.d.O., Adair, T., and Marinho, M.d.F. (2020). Estimating completeness of national and subnational death reporting in Brazil: Application of record linkage methods. *Population Health Metrics* 18(22). doi:10.1186/s12963-020-00223-2.

Dechter, A.R. and Preston, S.H. (1991). Age misreporting and its effects on adult mortality estimates in Latin America. *Population Bulletin of the United Nations* 31–32: 1–16.

Di Lego, V., Turra, C.M., and Cesar, C. (2017). Mortality selection among adults in Brazil: The survival advantage of Air Force officers. *Demographic Research* 37(41): 1339–1350. doi:10.4054/DemRes.2017.37.41.

Dowd, J.B. and Hamoudi, A. (2014). Is life expectancy really falling for groups of low socio-economic status? Lagged selection bias and artefactual trends in mortality. *International Journal of Epidemiology* 43(4): 983–988. doi:10.1093/ije/dyu120.

Glei, D.A., Barbieri, M., and Santamaría-Ulloa, C. (2019). Costa Rican mortality 1950–2013: An evaluation of data quality and trends compared with other countries. *Demographic Research* 40(29): 835–864. doi:10.4054/DemRes.2019.40.29.

Glei, D.A., Paz, A.B., Aburto, J.M., and Barbieri, M. (2021). Mexican mortality 1990–2016: Comparison of unadjusted and adjusted estimates. *Demographic Research* 44(30): 719–758. doi:10.4054/DemRes.2021.44.30.

Gomes, M.M.F. and Turra, C.M. (2009). The number of centenarians in Brazil: Indirect estimates based on death certificates. *Demographic Research* 20(20): 495–502. doi:10.4054/DemRes.2009.20.20.

Gonzaga, M.R., Lima, E.E.C., Queiroz, B.L., Ansiliero, G., and Freire, F.H.M.d.A. (2022). Mortality differentials in beneficiaries of the National Institute of Social Security of Brazil in 2015. *Revista Contabilidade & Finanças* 33(90). doi:10.1590/1808-057x20221556.en.

Gupta, A. and Mani, S.S. (2022). Assessing mortality registration in Kerala: The MARANAM study. *Genus* 78(1): 1–20. doi:10.1186/s41118-021-00149-z.

Hill, K. (1987). Estimating census and death registration completeness. *Asian and Pacific Population Forum/East-West Population Institute, East-West Center* 1(3): 8–13.

Hill, K. (1991). Approaches to the measurement of childhood mortality: A comparative review. *Population Index* 57(3): 368–382. doi:10.2307/3643873.

Hill, K., Choi, Y., and Timæus, I.M. (2005). Unconventional approaches to mortality estimation. *Demographic Research* 13(12): 281–300. doi:10.4054/DemRes.2005.13.12.

Hill, K., You, D., and Choi, Y. (2009). Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data errors. *Demographic Research* 21(9): 235–254. doi:10.4054/DemRes.2009.21.9.

IBGE (2003). *Censo Demográfico 2000: resultados de pesquisa de avaliação da cobertura da coleta*. Textos para discussão 9, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

IBGE (2018). *Projeções da população: Brasil e unidades da federação, Revisão 2018*. Tech. rep., Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

Jdanov, D.A., Jasilionis, D., Soroko, E.L., Rau, R., and Vaupel, J.W. (2008). Beyond the Kannisto–Thatcher database on old age mortality: An assessment of data quality at advanced ages. Unpublished Manuscript. doi:10.4054/MPIDR-WP-2008-013.

Kannisto, V., Jeune, B., and Vaupel, J. (1999). Assessing the information on age at death of old persons in national vital statistics. In: Jeune, B. and Vaupel, J.W. (eds.). *Validation of exceptional longevity*. Odense: Odense University Press: 235–249.

Li, N. and Gerland, P. (2013). Using census data to estimate old-age mortality for developing countries. Paper prepared for and presented to session 17-05: Indirect methods of mortality and fertility estimation: New techniques for new realities. Busan: XXVII IUSSP International Population Conference.

Martins, L.H. (2022). Estimativas indiretas de expectativa de vida em idades avançadas no Brasil e suas regiões. Mestrado em Demografia. Belo Horizonte: Universidade Federal de Minas Gerais.

Myers, R.J. (1940). Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America* 41(2): 395–415.

Nam, C.B. (1995). Another look at mortality crossovers. *Social Biology* 42(1–2): 133–142. doi:10.1080/19485565.1995.9988893.

Nepomuceno, M.R. and Turra, C.M. (2020). The population of centenarians in Brazil: Historical estimates from 1900 to 2000. *Population and Development Review* 46(4): 813–833. doi:10.1111/padr.12355.

Ouedraogo, S. (2020). Estimation of older adult mortality from imperfect data. *Demographic Research* 43(38): 1119–1154. doi:10.4054/DemRes.2020.43.38.

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. doi:10.1080/00324728.2021.1918752.

Palloni, A. and Pinto-Aguirre, G. (2011). Adult mortality in Latin America and the Caribbean. In: Rogers, R.G. and Crimmins, E.M. (eds.). *International handbook of adult mortality*. Dordrecht: Springer Netherlands: 101–132. doi:10.1007/978-90-481-9996-9_5.

Peralta, A., Benach, J., Borrell, C., Espinel-Flores, V., Cash-Gibson, L., Queiroz, B.L., and Marí-Dell'Olmo, M. (2019). Evaluation of the mortality registry in Ecuador (2001–2013) – social and geographical inequalities in completeness and quality. *Population Health Metrics* 17(3): 1–12. doi:10.1186/s12963-019-0183-y.

Preston, S.H., Elo, I.T., Rosenwaike, I., and Hill, M. (1996). African-American mortality at older ages: Results of a matching study. *Demography* 33(2): 193–209. doi:10.2307/2061872.

Preston, S.H., Elo, I.T., and Stewart, Q. (1999). Effects of age misreporting on mortality estimates at older ages. *Population Studies* 53(2): 165–177. doi:10.1080/00324720308075.

Pullum, T.W. (1991). Statistical methods to adjust for date and age misreporting to improve estimates of vital rates in Pakistan. *Statistics in Medicine* 10(2): 191–200. doi:10.1002/sim.4780100205.

Queiroz, B.L., Gonzaga, M.R., Vasconcelos, A., Lopes, B.T., and Abreu, D.M. (2020). Comparative analysis of completeness of death registration, adult mortality and life expectancy at birth in Brazil at the subnational level. *Population Health Metrics* 18(1): 1–15. doi:10.1186/s12963-020-00213-4.

Richman, R.D. (2017). Old age mortality in South Africa, 1985–2011. Master's thesis. Cape Town: University of Cape Town.

Romero Prieto, J., Verhulst, A., and Guillot, M. (2021). Estimating the infant mortality rate from DHS birth histories in the presence of age heaping. *PLoS One* 16(11): e0259304. doi:10.1371/journal.pone.0259304.

Schmertmann, C.P. and Gonzaga, M.R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography* 55(4): 1363–1388. doi:10.1007/s13524-018-0695-2.

Spoorenberg, T. and Dutreuilh, C. (2007). Quality of age reporting: Extension and application of the modified Whipple's index. *Population* 62(4): 729–741. doi:10.3917/pope.704.0729.

Thatcher, A.R., Kannisto, V., and Vaupel, J.W. (1998). *The force of mortality at ages 80 to 120*. Odense: Syddansk Universitetsforlag.

Turra, C.M., Fernandes, F., Almeida Calazans, J., and Nepomuceno, M.R. (2023). Age reporting for the oldest old in the Brazilian COVID-19 vaccination database: What can we learn from it? *Demographic Research* 48(28): 829–848. doi:10.4054/DemRes.2023.48.28.

Whipple, G.C. (1919). *Vital statistics: An introduction to the science of demography*. Hoboken: John Wiley & sons, Incorporated.

Appendix

A-1 Change in e_x with a change in mortality at a single-year age

Cumulative mortality at exact age x is

$$H(x) = \int_0^x \mu(a)\, da\,. \tag{27}$$

When mortality rates are a step function with values $\mu_0, \mu_1, \ldots$ over discrete, non-overlapping, single-year age intervals $A_0, A_1, \ldots$, then cumulative mortality through age x is

$$H(x) = \int_0^x \left(\sum_i I(a \in A_i)\, \mu_i \right) da\,, \tag{28}$$

and its derivative with respect to a particular mortality rate μ_y, where y is an integer, is

$$\frac{\partial H(x)}{\partial \mu_y} = \int_0^x I(a \in A_y)\, da\,. \tag{29}$$

From here on we use subscripts rather than functional notation when exact ages x and y are integers. Because $H_x = -\ln \ell_x$

$$\begin{aligned} \frac{\partial \ell_x}{\partial \mu_y} &= -\ell_x \frac{\partial H_x}{\partial \mu_y} \\ &= -\ell_x \int_0^x I(a \in A_y)\, da \\ &= -I(x > y) \cdot \ell_x\,. \end{aligned} \tag{30}$$

The derivative of person-years lived after exact age x with respect to the mortality rate in $A_y = (y, y+1]$ is

$$\begin{aligned}
\frac{\partial T_x}{\partial \mu_y} &= -\int_{a=x}^{\infty} \ell_a \int_{z=0}^{a} I(z \in A_y)\, dz\, da \\
&= -\int_{z=0}^{\infty} I(z \in A_y) \int_{a=max(x,z)}^{\infty} \ell_a\, da\, dz \\
&= -\int_{z=0}^{\infty} I(z \in A_y)\, T_{max(x,z)}\, dz \\
&= -\int_{y}^{y+1} T_{max(x,z)}\, dz \\
&= -I(x \le y) \cdot \left[\int_{y}^{y+1} T(z)\, dz\right] - I(x > y) \cdot T_x \\
&= -I(x \le y) \cdot \bar{T}_y - I(x > y) \cdot T_x\,,
\end{aligned} \tag{31}$$

where $\bar{T}_y = \frac{1}{2}[T_y + T_{y+1}]$ is the standard trapezoidal approximation to $\int_y^{y+1} T(z)dz$.

Together these results imply

$$\begin{aligned}
\frac{\partial e_x}{\partial \mu_y} &= \frac{\ell_x \frac{\partial T_x}{\partial \mu_y} - T_x \frac{\partial \ell_x}{\partial \mu_y}}{\ell_x^2} \\
&= \tfrac{1}{\ell_x}\left(\tfrac{\partial T_x}{\partial \mu_y} - e_x \tfrac{\partial \ell_x}{\partial \mu_y}\right) \\
&= -I(x \le y) \cdot \frac{\bar{T}_y}{\ell_x} - I(x > y) \cdot e_x + e_x \cdot I(x > y) \\
&= -I(x \le y) \frac{\bar{T}_y}{\ell_x}\,.
\end{aligned} \tag{32}$$

An important special case of Equation (32), which we use extensively in the main text of the article, is for life expectancy at birth ($x = 0$):

$$\frac{\partial e_0}{\partial \mu_y} = -\bar{T}_y\,. \tag{33}$$

A-2 Derivative of $T(a)\mu(a)$ over age

The behavior of the life table function $T(a)\cdot\mu(a)$ over adult ages turns out to be important for determining how census and death registration errors affect life expectancy estimates. In particular, it is essential to understand whether this function increases or decreases with age.

At ages over which the log mortality rate increases at a rate b over age, the derivative of $T(a)\cdot\mu(a)$ with respect to age a is

$$\begin{aligned}\frac{\partial T(a)\mu(a)}{\partial a} &= T'(a)\mu(a) + T(a)\mu(a)' \\ &= -\ell(a)\mu(a) + T(a)\mu(a)\tfrac{\partial \ln \mu(a)}{\partial a} \\ &= -\ell(a)\mu(a)\left[1 - e(a)\cdot b\right] \\ &= b\cdot d(a)\left[e(a) - \tfrac{1}{b}\right] .\end{aligned} \tag{34}$$

Because $b \approx .10$ at high adult ages in human populations, this result implies that $T(a)\mu(a)$ will tend to increase at ages for which remaining life expectancy is above approximately 10 years, reach a maximum when $e_a \approx 10$, and decrease at advanced ages for which remaining life expectancy is less than 10 years. The rate of change in $T(a)\mu(a)$ will also tend to be higher, everything else equal, at ages for which we expect more life table deaths $d(a)$.

A-3 Converting from single-year to age-group age misstatement matrices

Suppose that the vector of the true population by single-year age is $\eta \in \mathbb{R}^A$and the matrix of age misstatement is $\{p_{xy}\}$, where row x is the reported age, and column y is the true age. The complete $A \times A$ matrix of exposure by (reported age, true age) is then

$$N_1 = P\,\mathrm{diag}(\eta)\,. \tag{35}$$

Define an exhaustive and mutually exclusive set of age groups $g = 1 \ldots G$ and a $G \times A$ matrix of 0s and 1s: $W = \{w_{gy} = \mathbf{1}[\text{age } y \text{ in group } g]\}$, which tells us whether age group g contains single-year age y. The complete $G \times G$ matrix of exposure by (reported age group, true age group) is then

$$N_{group} = W N_1 W' = W\,P\,\mathrm{diag}(\eta)\,W'\,. \tag{36}$$

With this notation the number of people with true group Y who are reported in group X is

$$n_{XY} = w'_X P \,\mathrm{diag}(\eta)\, w_Y \;, \tag{37}$$

where w'_X is the xth row of W.

A-4 A parametric model for P

A-4.1 Model definition

We use a six-parameter model for the single-year age misstatement matrix P. At each age $y \in \{0, 1, \ldots (A-1)\}$ there are probabilities of correct age reporting (P_C), age understatement (P_U), and age overstatement (P_O) that must sum to one. We describe these probabilities with a multinomial logit model in which

$$[P_C,\, P_U,\, P_O]_y \;\propto\; [1,\, e^{\alpha_U + \beta_U\, y},\, e^{\alpha_O + \beta_O\, y}] \,. \tag{38}$$

Parameters $(\alpha_U, \beta_U, \alpha_O, \beta_O)$ determine the probabilities of age underunderstatement, correct age reporting, and age overstatement for each true age. The probabilities of specific reported ages x, given true age y, are

$$p_{xy} \;=\; \begin{cases} P_U(y)\, \frac{\rho_U^{|x-y|}}{\sum_{x<y} \rho_U^{|x-y|}} & \text{for } x < y \\ P_C(y) & \text{for } x = y \\ P_O(y)\, \frac{\rho_O^{|x-y|}}{\sum_{x>y} \rho_O^{|x-y|}} & \text{for } x > y \,, \end{cases} \tag{39}$$

where the ρ parameters lie in the unit interval and represent the rate of geometric decay in the probability of misreports as we move from true age y to ages x that are increasingly distant. For example $\rho_U = 0.3$ means that understatement by 2 years is 30% as likely as understatement by 1 year, that understatement by 3 years is 30% as likely as understatement by 2 years, and so forth.

A-4.2 Parameter estimation

For any specific value of $\theta = (\alpha_U, \beta_U, \alpha_O, \beta_O, \rho_U, \rho_O)$ we calculate the $A \times A$ misstatement matrix $P(\theta)$ and the set of expected age-group counts $\hat{n}_{XY}$ from Equation (37). We

then choose θ to minimize the sum of squared differences between the expected group counts and those predicted by a published group misstatement matrix, $n_{XY}^{\text{published}}$:

$$\hat{\theta} = \text{argmin} \sum_X \sum_Y [\ \hat{n}_{XY}(\theta) - n_{XY}^{\text{published}}\ ]^2 \,, \tag{40}$$

which in our model is

$$\hat{\theta} = \text{argmin} \sum_X \sum_Y [\ w'_X\, P(\theta)\, \text{diag}\,(\eta)\, w_Y - n_{XY}^{\text{published}}\ ]^2 \,. \tag{41}$$

Notice that Equation (41) requires single-year counts (denoted η for exposure in the equation, but they could also be deaths δ) from a reference population. For the African American pattern we fit Equation (41) using São Paulo stationary deaths with the age misstatement information from Preston, Elo, and Stewart (1999: Table 2). For the India pattern we used the São Paulo stationary population with group error probabilities from Bhat (1990: Table 3: Males). Best fitting parameters are shown in Table A-1.

Table A-1: Model parameters that minimize the objective function in Equation (41), for alternative five-year age misreporting patterns

	α_U	β_U	α_O	β_O	ρ_U	ρ_O
African American	–4.449	0.046	–20.901	0.226	0.820	0.212
India	–2.946	0.150	–2.209	0.149	0.774	0.764