Estimation of Life Tables in the Latin American Data Base (LAMBdA): Adjustments for Relative Completeness and Age Misreporting

Alberto Palloni* Guido Pinto† Hiram Beltrán-Sánchez‡

October 27, 2016

1 Background

The methodologies described in this paper belong to a small subset of a broader set of methods developed to produce adjusted estimates of adult mortality for countries in the Latin American and Caribbean (LAC) region covering 150-160 years, from 1850 to 2010. This period encompasses approximately the end of colonial rule, the aftermath of wars of independence from Spanish and Portuguese domination, the establishment of nation states, integration into a world system and the world economy, and all developments that unfolded following World War II.[1] In this paper we focus only on adjustments of life tables for the post-1950 period. To do so we avail ourselves of mortality data consisting of yearly deaths by age and gender and population censuses. Because methods to adjust for completeness of death registration are well-known we focus on the description of relatively new methods to adjust for adult age misreporting. We then combine these two methods in an evaluation study designed to identify an optimal strategy to construct adjusted life tables for adult ages.

The paper is in six sections. In the first section we briefly define problems caused by defective vital statistics and census enumerations. In the second section we propose a model to represent the nature of adult age misreporting and in the third section we describe a methodology to detect and adjust for adult age misreporting. The fourth section describes an evaluation study designed to assess the performance of techniques to correct mortality indicators for both errors of coverage and age reporting. The fifth section discusses results from the evaluation study. The last section summarizes results and argues that adjustment of imperfect mortality data is subject to uncertainty and that treatment of the adjusted data is best carried out with models that account for uncertainty.

2 Errors affecting measures of adult mortality

The post-1950 mortality data in LAC is limited by defective coverage and adult age misreporting. By and large, observed death counts are a variable fraction of the 'true' number of deaths that take

*Center for Demography and Health of Aging, Center for Demography and Ecology, University of Wisconsin-Madison. Email: palloni@ssc.wisc.edu.

†Center for Demography and Health of Aging, Center for Demography and Ecology, University of Wisconsin-Madison.

‡Department of Community Health Sciences & California Center for Population Research, Fielding School of Public Health, UCLA. Email: beltrans@ucla.edu.

[1] We define as "adult" the population aged 5 and older and as "children" those younger than age 5.

place at a particular time as they exclude events that, for a number of reasons, are never recorded. Since population censuses too are normally affected by coverage problems, mortality rates computed with the raw data may contain smaller *net errors* that would be expected otherwise. In general, however, the observed mortality rates underestimate mortality levels, particularly at very young and old ages. We use the term *relative completeness* when we speak of ratios of observed to true mortality rates.

Table 1 displays estimates of relative completeness of adult (over 5 years of age) and, for comparison, those corresponding to infant (age 0) and early child (ages 1-4) death registration in a sample of LAC countries over two different periods of time. The figures in this table confirm that the quality of the information is poorer at very young ages and that, although there is a clear universal trend toward improvement, an important fraction of countries still show signs of deficient registration even quite recently.

Imperfect relative completeness of death registration is not the only problem affecting estimates of mortality. An important domain of errors involves age misreporting and the most insidious manifestation is systematic over (under) reporting. Vital and census statistics in LAC countries are, almost without exception, affected by age overstatement, particularly at ages over 40 or 45 (see below). When the (true) age distribution of a population is roughly exponential in nature —as it always is in stable and quasi stable populations—systematic age overstatement of populations induces downward biases in mortality rates at older ages. These biases are not offset when there is an equal propensity to overstate ages at death. The reason these two type of errors do not cancel each other out is that while both adult mortality rates and adult population age distributions are roughly exponential, one slopes upwards (mortality rates) whereas the other slopes downwards (population). Matters are made worse when, as is almost always the case, the rate of decrease of population with age (natural rate of increase in a stable population) is several times lower than the rate of increase of adult mortality rates (rate of senescence in Gompertz mortality regimes). The consequence is that unless the propensity to overestimate ages at death is much higher than the propensity to overestimate ages of population, observed mortality rates will contain downward biases. If left uncorrected, the resulting life tables will offer a misleading portrayal of the curvature of mortality at older ages, suggesting the existence of slower rates of senescence or heavy influence of selection due to changing frailty composition. As the quality of vital registration and census enumeration improves, the magnitude of these biases tends to decrease and the entire history of observed life tables will erroneously suggest trends in old age patterns of mortality and even relative acceleration of the rates of mortality decline at older ages.

Unlike problems created by age heaping, distortions caused by systematic age misstatement cannot be repaired by restoring the original age distribution standard using computations that rely on safe assumptions. Systematic age misstatement is altogether different since it is harder to diagnose and, as we show below, its treatment requires additional knowledge of two functions: (a) the conditional (on age and gender) propensity of individuals to exaggerate (decrease) the true age and (b) the conditional (on age and gender) distribution of the difference between the correct and declared age. To solve the problem we propose generalizations of an existing procedure to identify the presence of age misstatement, formulate a new method to estimate functions describing (a) and (b) from observables, and define an algorithm that adjusts observed adult mortality rates for both faulty coverage and systematic age misreporting.

Table 2 displays estimated biases in mortality rates at ages over 45 in a sample of country-years used in our analysis and the corresponding errors in life expectancy at age 60.

The problems generated by defective completeness of death registration as well as alternative

Table 1: Relative completeness of deaths registration in the LAC countries: 1920-2010.

Country	Period 1900-1949				Period 1950+	
	Mid-Year	Age 0	Age 1-4	Age 5+	Mid-Year	Age 5+
Argentina	1914	0.968	0.865	0.939	1953	0.974
					2005	0.995
Brazil					1985	0.885
					2005	0.996
Chile	1925	0.867	0.829	0.852	1956	0.961
	1945	0.867	0.829	0.934	2006	0.980
Colombia	1944	0.821	0.815	0.749	1957	0.790
					2008	0.800
Costa Rica	1927	0.901	0.922	0.893	1956	0.918
	1938	0.901	0.922	0.893	2005	0.975
Cuba	1925	0.806	0.893	0.800	1961	0.890
	1948	0.806	0.893	0.870	2006	0.989
Dominican Republic	1942	0.476	0.451	0.487	1955	0.500
					2006	0.604
Ecuador					1956	0.738
					2005	0.805
El Salvador	1940	0.554	0.776	0.721	1955	0.700
					2008	0.714
Guatemala	1945	0.714	0.898	0.784	1957	0.888
					2005	0.940
Honduras	1942	0.542	0.551	0.495	1955	0.518
	1947	0.542	0.551	0.500	1989	0.750
Mexico	1925	0.843	0.822	0.752	1955	0.860
	1945	0.843	0.822	0.883	2005	0.959
Nicaragua	1945	0.526	0.545	0.498	1956	0.456
					2007	0.561
Panama	1945	0.837	0.757	0.829	1955	0.839
					2005	0.853
Paraguay					1956	0.601
					2006	0.681
Peru					1950	0.490
					2008	0.533
Uruguay	1908	0.844	0.822	0.879	1969	0.960
					2007	0.996
Venezuela	1938	0.833	0.857	0.846	1955	0.866
	1945	0.833	0.857	0.855	2006	0.895

Table 2: Biases due to age overstatement.

Country	Mid-Year	Unadjusted		Adjusted*	
		E(45)	E(60)	E(45)	E(60)
Argentina	1953	25.96	15.39	25.29	14.55
	2005	30.02	17.96	29.33	17.15
Brazil	1985	28.55	17.61	27.62	16.51
	2005	31.27	19.77	30.23	18.58
Chile	1956	24.44	14.57	23.72	13.64
	2006	33.20	20.45	32.16	19.33
Colombia	1957	27.34	16.68	26.46	15.67
	2008	35.09	22.29	33.86	20.96
Costa Rica	1956	29.08	17.55	28.10	16.46
	2005	34.96	22.40	33.78	21.13
Cuba	1961	30.13	18.15	29.18	17.08
	2006	33.46	20.94	32.56	19.95
Dominican Republic	1955	33.62	22.44	31.91	20.52
	2006	38.35	25.76	36.41	23.68
Ecuador	1956	28.75	17.98	27.77	16.83
	2005	37.42	25.23	35.94	23.62
El Salvador	1955	27.64	17.54	26.69	16.42
	2008	32.79	21.74	31.85	20.62
Guatemala	1957	24.44	15.06	23.68	14.07
	2005	31.39	20.22	30.42	19.10
Honduras	1955	30.55	20.37	29.14	18.64
	1989	37.33	25.06	35.61	23.17
Mexico	1955	26.57	16.69	25.80	15.71
	2005	33.04	21.13	31.97	19.95
Nicaragua	1956	32.09	21.05	30.61	19.37
	2007	36.23	24.05	34.71	22.41
Panama	1955	28.93	17.67	27.87	16.45
	2005	35.92	23.18	34.65	21.81
Paraguay	1956	32.97	20.81	31.73	19.44
	2006	34.84	22.17	33.60	20.84
Peru	1950	30.61	20.64	29.47	19.25
	2008	39.37	26.32	37.66	24.52
Uruguay	1969	26.72	15.47	26.11	14.69
	2007	30.35	18.17	29.85	17.57
Venezuela	1955	27.49	16.81	26.47	15.64
	2006	32.75	20.94	31.53	19.59

* Adjusted for age misreporting

adjustments procedure to deal with it are well-known. Much less is known about the nature and impact of age misreporting. In the section below we propose a methodology to identify the presence of these errors and to correct them.

3 Systematic age misreporting

3.1 Setup

We begin with a few basic definitions. Let θ_x^o be the average conditional probability that individuals aged x overstate their age in a census and θ_x^u the conditional probability of understating their age. Then $(1-\theta_x^o-\theta_x^u)$ is the probability of an accurate age statement. Individuals who over(under) state their age do so by choosing, not always randomly, the age declared and observed in the census. This age could be $n>0$ years removed from the true age. As we show below, it suffices to let n range between 1 and 10+ since the frequencies for values of 10 years and above are exceedingly small, e.g. individuals rarely over(understate) their age by more that ten digits. Let $\rho_x^o(n)$ be the average conditional probability that individuals aged x who overstate ages will do so by n years with an analogous definition for the probabilities for age understatement, $\rho_x^u(n)$ and with $\sum_n \rho_x^u(n)=\sum_n \rho_x^o(n)=1$. To compute the observed number at age y, P_y^o, we consider the true number at that age P_y^T, and apply the conditional probabilities defined above:

$$P_y^o = P_y^T(1-\theta_x^o-\theta_x^u) + \sum_{j=1}^{j=10} P_{y-j}^T \rho_{y-j}^o(j)\theta_{y-j}^o + \sum_{j=1}^{j=10} P_{y+j}^T \rho_{y+j}^u(j)\theta_{y+j}^u \ . \tag{3.1}$$

This expression can be generalized for all ages between 0 and 100 in compact matrix notation:

$$\Pi^o = \Theta\Pi^T \tag{3.2}$$

where Π^o is the (101x1) observed population vector, Π^T is the (101x1) true population vector and Θ is a 101x101 square matrix of "transition" probabilities, e.g. the probabilities of migration into or out of single year age-groups. In particular, the diagonal of Θ contains the probabilities of correctly declaring ages, $(1-\theta_x^o-\theta_x^u)$, and entries in the off-diagonal row k for columns $k-1, k-2, \ldots, k-10$ are the values $\rho_{y-1}^o(j)\theta_{y-1}^o, \ldots, \rho_{y-10}^o(j)\theta_{y-10}^o$ whereas those in columns $k+1, k+2, \ldots, k+10$ are the values $\rho_{y+1}^u(j)\theta_{y+1}^u, \ldots, \rho_{y+10}^u(j)\theta_{y+10}^u$. One can retrieve the matrix with the true age distribution of the population after pre-multiplying the previous expression by the inverse of Θ^{-1}, that is

$$\Theta^{-1}\Pi^o = \Pi^T \ , \tag{3.3}$$

an operation that requires full knowledge of the matrix Θ. As we show below, demographers have only superficial information about the nature of this matrix in LAC countries or anywhere else for that matter (but see Bhat (1990)). In the absence of precise knowledge of the probabilities contained in the matrix one could adopt shortcuts, simplifications that circumvent knowledge gaps but that, as shown below, lead to identification problems, most of which translate into inability to specify an invertible matrix of transition probabilities.

3.2 Observed patterns of age misreporting

What do we know about age misreporting in population and death counts in LAC and in other countries? There is an extensive literature on general errors in age reporting (Ewbank, 1981; Chidambaram and Sathar, 1984; Kamps E., 1976; Nuñez, 1984) as well as on systematic age misstatement, mostly adult age overstatement, in population counts. And while a fair number of these studies uncover evidence of overstatement in low income countries (Mazess and Forman, 1979; Grushka, 1996; Bhat, 1987, 1990; Del Popolo, 2000; Dechter and Preston, 1991) or in US migrant (Hispanic or Hispanic origins) groups (Rosenwaike and Preston, 1984; Spencer, 1984), there is a body of literature that identifies patterns of age overstatemet in high income countries as well (Horiuchi and Coale, 1985; Coale and Kisker, 1986; Condran et al., 1991; Preston et al., 2003; Elo and Preston, 1994). In the US, for example, age overstatement is one of the factors that could explain the so called Black-White mortality crossover, whereby African American mortality rates dip below those of their White counterparts at very old ages (over 70). And while the recurrent idea of heavy selection due to frailty has not been completely discarded, the most recent investigations suggest that overstatement of ages in the population (and also deaths) among African American more so than among Whites accounts for a substantial part of the mortality crossover (Elo and Preston, 1994). The Black-White mortality crossover is just an extreme example of the damage that age misreporting can inflict on estimates of adult mortality. As others before us have done (Dechter and Preston, 1991; Grushka, 1996; Bhat, 1987, 1990), we will show that age overstatement is also an important source of error in LAC countries.

Partial information on the matrix Θ has been obtained mostly from studies involving record linkages (Elo and Preston, 1994; Preston et al., 1996; Rosenwaike and Preston, 1984; Rosenwaike, 1987), post enumeration surveys (Ortega and Garcia, 1985) and comparisons of two independently gathered data sources that should produce the same outcomes (Bhat, 1990). In all these studies, however, the information is either aggregated in five-year age groups or applies to populations with levels of education that are much higher than those in LAC countries. Lack of age detail is problematic since computation of conditional probabilities in coarse age groups rests on approximations that, if violated, are generally harmful to the accuracy of estimates. Using a transition matrix appropriate for a population with higher or lower levels of education or literacy than the target one may lead to distortions since age misstatement is strongly associated with levels of education.

3.3 Misreporting of ages of population

To circumvent the foregoing problems we take advantage of a 2002 evaluation study launched by the Central American Center for Population at the University of Costa Rica. The program was designed to assess the quality of information of death registration and the accuracy of the 2000 census counts[2]. One of the components of this study was a linkage of an age stratified sample of 9,113 individual census records with the national voter registers, a database that contains age information from birth certificates. A total of 7426 records were matched corresponding to 81.5% of the original sample and 86.6 % of the non foreign born part of the sample. The final data set contains individuals classified by gender, education and other traits, and by 'true' and declared age. To estimate the entries of matrix Θ we proceeded in two steps:

i *Estimation of probabilities of age over and understatement, $\theta_x^o(V)$ and $\theta_x^u(V)$ where V is a vector of individual characteristics, including age*: We first estimate a logistic model for a

[2] We are grateful to Drs. Gilbert Brenes and Luis Rosero Bixby from the Central American Population Center at the University of Costa Rica for having provided tabulations we used in this study.

binary variable set to 1 when there is over (under) statement and zero otherwise. Initially the model specifies a vector of covariates including age, age squared, urban/rural residence, gender, and education. The sample includes individuals aged 50 and over since at younger ages there are only traces of systematic age misstatement (mostly in the form of heaping). Because gender and age are the only covariates that can be used at a national level, we simplify the model to include only these two traits as predictors. Finally, after verifying that the effects of age squared and gender were statistically insignificant, the final model conditions only on 'true' age of individuals. Table 3 displays estimated parameters for over and under stating ages using the weighted sample.

ii *Estimation of conditional probabilities of over(under) stating ages by* $1 < n \leq 10$ *years,* $\rho_x^o(j)$ *and* $\rho_x^u(j)$: We estimate a multinomial model with 9 categories that includes gender and (true) continuous age as independent variable. The resulting estimates reveal that the effects of gender are always statistically insignificant, that those of age show no clear pattern and, in addition, that their magnitude is quite small in 6 out of 8 cases for overstatement models and in 5 out of 8 contrasts for age understatement. To simplify we estimate a null model predicting the average conditional probabilities of exaggerating (or diminishing) by n years applicable to all ages older than 50 and both genders. The values of the predicted probabilities of over and understating the true age are in Table 4.

Although it is now possible to compute an estimator of the target mobility matrix, $\hat{\Theta}$, there remains a knotty problem of identification that cannot be resolved without additional simplifications. Suppose, for example, we seek to estimate mortality trends in a country with much lower levels of education than in Costa Rica. Replacing $\hat{\Theta}$ for the true matrix in (3.3), we will obtain a true distribution of ages but only under the very strong assumption that age misstatement is identical across countries. This contradicts accumulated knowledge showing that the severity of age misstatement increases as levels of education drop. A less constraining assumption is to argue that while the *age pattern* of misstatement is identical across countries, the levels could be different. To express this one could think of multiplying the conditional probabilities of over and under stating ages (or a monotonic transform of it) by some constant, say ϕ^o and ϕ^u for over and understatement respectively. While this is a reasonable strategy it generates an additional problem, namely, that a unique solution for equation (3.2) may no longer be possible since different combinations of ϕ^o and ϕ^u embedded in the transition matrix could plausibly yield identical results. To circumvent this new difficulty we propose a standard pattern of *probabilities of net age overstatement* as $\varphi_x^S = \theta_x^o - \theta_x^u$ and then apply to it the conditional probabilities of overstating one's age by n years (the $\rho_x^o(j)$ values defined before). Under these conditions the off-diagonal cells of the matrix defined by φ_x^S, $\hat{\Theta}^S$, simplify as all entries involving age understatement become zeros. This makes identification more likely and the search for a unique solution of ϕ^{no}, a parameter measuring the magnitude of the net overstatement (no) relative to the standard pattern, a more feasible enterprise.

There are two conditions required for this standard pattern to play a helpful role. The first is that the probabilities of age overstatement always be larger than the probabilities of age understatement. The second is that the conditional distribution of n, the integer number of years by which individuals exaggerate (diminish) their true age, be approximately the same among those who over and understate ages. Figure 1 displays predicted probabilities of over and understating ages by age, $\theta_x^o - \theta_x^u$, Figure 2 displays the differences $\varphi_x^S = \theta_x^o - \theta_x^u$, and Figure 3 shows predicted conditional probabilities of over stating ages by n years with $0 < n \leq 10$ or $\rho_x^o(j)$. These figures show that the first condition is always satisfied whereas the second is only approximately met in

Table 3: Estimated parameters of best logistic models for age misreporting.

Variable	Overreporting Coeff(se)	Underreporting Coeff(se)
True age[1]	0.014(.0036)	0.002(.0040)
Constant	-2.127(.271)	-1.846(.297)
N	6290	6290

[1] Regressions estimated using sampling weights. Sample includes population with true age 60 and older and excludes ambiguous cases and foreign citizens.

Table 4: Average (conditional) probabilities of overreporting ages.

	Probability[1]	
n	Overstating	Understating
1	0.621	0.510
2	0.191	0.128
3	0.079	0.091
4	0.040	0.052
5	0.023	0.041
6	0.015	0.035
7	0.009	0.028
8	0.007	0.026
9	0.005	0.013
10+	0.009	0.060

[1]Predicted values computed from a null multinomial logistic model with 10 categories, n=1786 (males and females). Estimation using sampling weights. Figures may not add up to 1 due to rounding errors.

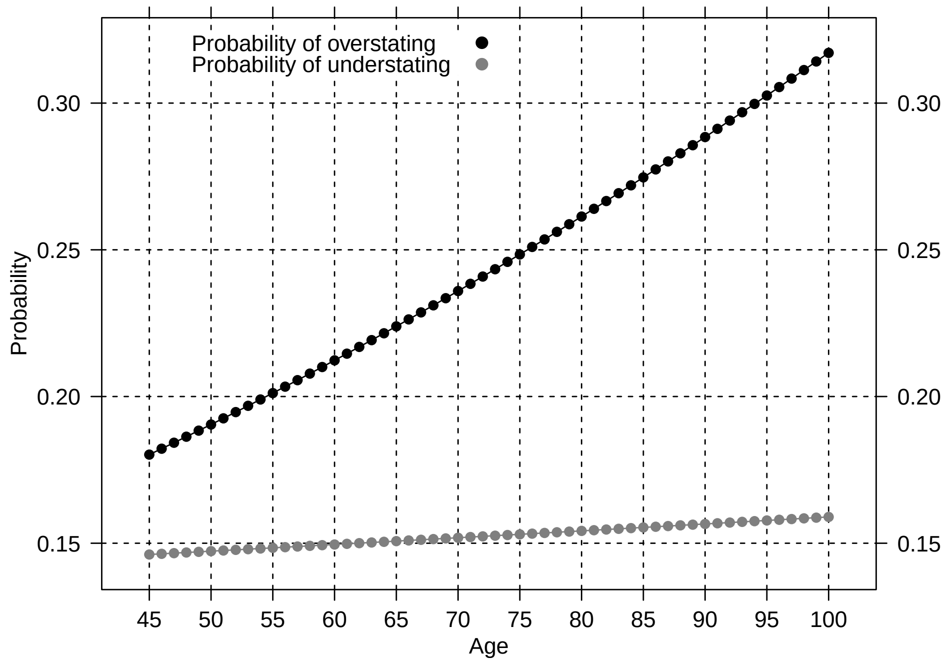
these data. However, differences are minor and are found mostly at higher values of n, where the probabilities of over(under) stating are small. We define these two items, the pair of age-specific differences between predicted probabilities of over and under statement (Table 3) and the associated conditional probabilities of overstating by n years (Table 4), to be the *standard pattern of age net overstatement*[3].

3.4 Misreporting of ages at deaths

The developments above only refer to age misreporting in population counts. However, it is known that mortality rates are also influenced by age misreporting of ages at death (Rosenwaike, 1987). The nature of the problem in this case is somewhat different since it is not the decedent that declares the age at death but a kin or someone else unrelated to the decedent. A handful of studies

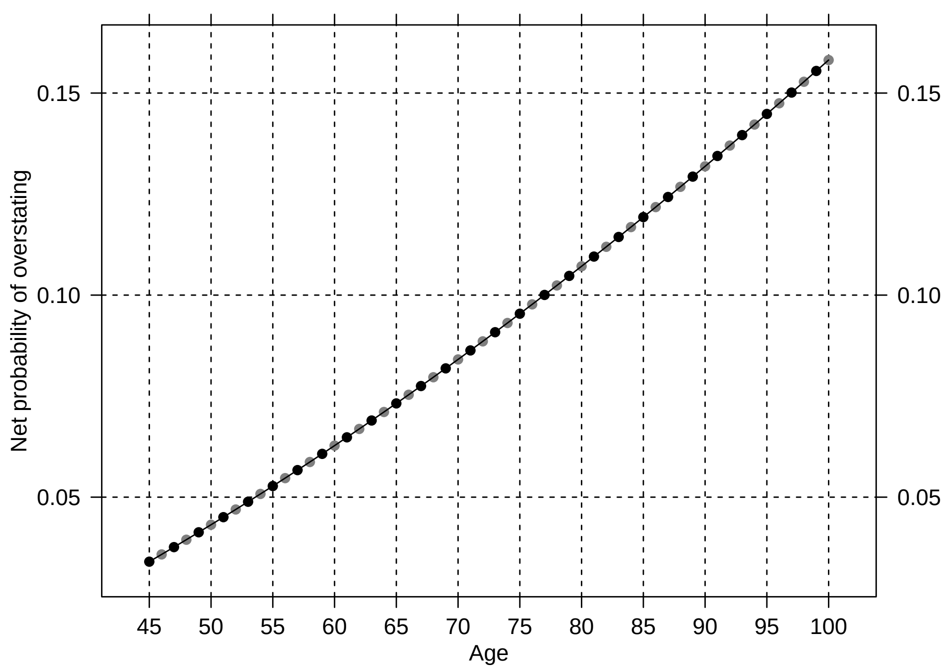
[3]The representation we used throughout suggests that patterns of age misreporting in any country is a multiple of the standard pattern. Although this helps the algebra and statement of proofs, we cheat in our computations and follow a roundabout algorithm. In fact, we generate new patterns of values from the standard by defining the function $\texttt{logit}(\varphi_x^i) = \alpha + \beta\texttt{logit}(\varphi_x^S)$, set the value of β equal to 1, and then identify the level of age overstatement in a population i by fixing α so that $\varphi_x^i \sim \phi^o \varphi_x^S$ and ϕ^o is the desired level of age over reporting.

Figure 1: Predicted probabilities of over(under) stating ages.



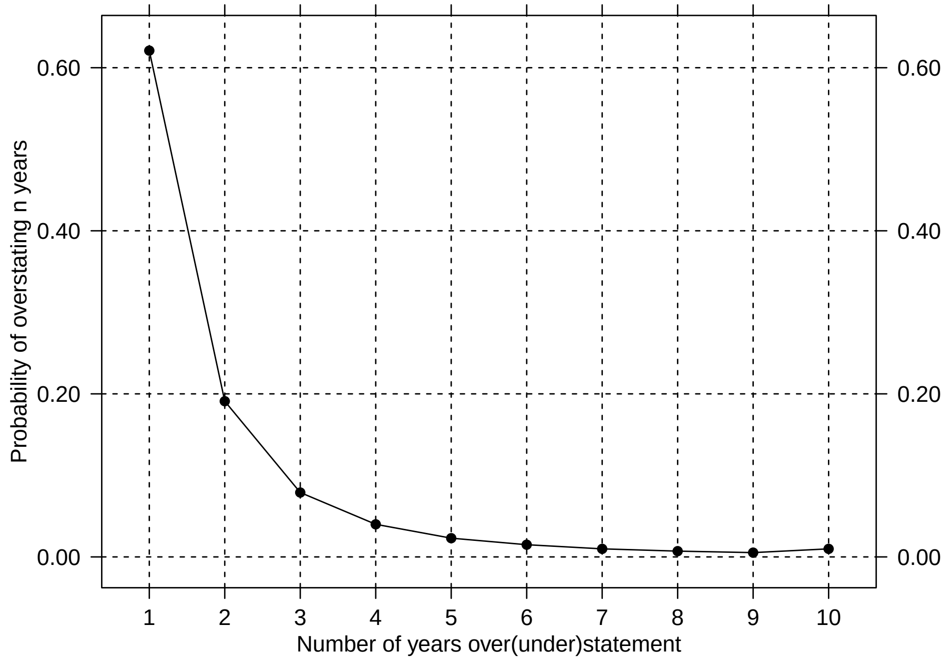
Source: Costa Rica Special study of 2000 population census.

Figure 2: Predicted probabilities of net overstating ages.



Source: Costa Rica Special study of 2000 population census.

Figure 3: Conditional probabilities of overstating age by n years.



Source: Costa Rica Special study of 2000 population census.

based on record linkages show that there is age misreporting of ages at death as well, albeit of lower magnitude than that found in population counts, and that it also tends to be in the direction of overstatement (Rosenwaike and Preston, 1984). This is confirmed by the application of indirect techniques designed to detect age at death overstatement in a number of low and high income countries (see below). It follows that expressions analogous to (3.1) and (3.2) must be applicable for death counts as well. To make the problem tractable one needs an empirical approximation to a matrix analogous to Θ but now specialized to ages at death. To our knowledge no such matrix has ever been estimated in LAC or anywhere else and we are unaware of any national data that could be used for such purpose. In what follows we assume that the *standard age pattern of age misstatement of death counts is identical to that of age misstatement of population counts*, although its level may be different. This assumption enables us to define the final model of age misreporting as a set of two equations with two unknown parameters:

$$\Pi^o = \phi^{no}\hat{\Theta}^S\Pi^T \tag{3.4}$$
$$\Delta^o = \lambda^{no}\hat{\Theta}^S\Delta^T \tag{3.5}$$

where $\Delta^T$and Δ^O are the true and observed distributions of death counts and λ^{no} is the magnitude of net overstatement of ages at death relative to the standard pattern. In closed populations equations (3.4) and (3.5) are naturally (see below) related and it is unlikely that there is always a unique solutions for ϕ^{no} and λ^{no} unless we either fix the value of one of them or, alternatively, retrieve solely their ratio. A brief proof of lack of identification is in Appendix B and solutions for empirical estimation are in section 4.2.

4 Identification and correction of errors due to systematic age misreporting

In this section we propose a methodology to identify and then adjust mortality statistics for age misreporting. The methodology is only applicable when age misreporting is produced following the model outlined in the previous section.

4.1 Identification of systematic age misreporting

A key component of our analysis is the detection and identification of patterns of age misstatement in the population and death counts. As shown in a previous section, the distortions associated with age misreporting in population and death counts is more complex than those involving only faulty completeness. Detection of the problem is difficult since its manifestations are quite subtle and, in the absence of overt and striking phenomena such as the US Black-White cross over, is likely to remain concealed and undetected. There are two well-tested methods to identify the existence of age over(under) statement in either population or death counts. The first method requires an external data source with correct dates of birth or ages in a population at a particular time that can be compared to age-specific census counts at approximately the same time. An example of this is the utilization of Medicare data in the US, a source of information that, as a rule, contains both population exposed and mortality data. Because Medicare data are linked to Social Security records and these are known to register age with high precision, mortality rates computed from Medicare data are a gold standard against which conventional mortality rates could be contrasted and their quality evaluated (Elo et al, 2004). If one ignores the existence of a population not covered by Medicare records, it is also feasible to link individual census records to Medicare records and investigate more precisely the nature of patterns of age misreporting in census counts. If, in addition, Medicare records are linked to the US National Death Index (NDI) it is then possible to repeat the same operations and assess the quality of reporting of age at deaths. In all cases one must assume that the coverage of population in both sources is complete or, if incomplete, identical[4]. Record linkage from multiple sources such as those illustrated above has rarely been used as it is expensive and involves resolution of complicated confidentiality issues.

A second method is less data demanding, considerably less expensive and is simple to apply but can only *reveal* the existence of age misreporting in one of the two sources and provides few clues about its nature. The procedure was proposed by Preston and colleagues (Rosenwaike and Preston, 1984; Elo and Preston, 1994; Bhat, 1990; Grushka, 1996) and has been applied in countries of North America, Western Europe and in Latin America (Condran et al., 1991; Grushka, 1996; Dechter and Preston, 1991; Palloni and Pinto, 2004; Del Popolo, 2000). In a nutshell the method consists of comparing cumulative population counts in a census in year t_1 to the expected cumulative population counts in a second population census in year t_2. The computation of expected quantities requires both an initial census opening the intercensal interval, a second census counts at time t_2 closing the intercensal interval, and age specific deaths counts in the intercensal period spanning an interval of $k = (t_2 - t_1 + 1)$ years. The ratio of observed to expected population is an indicator of age misstatement:

$$cmR^o_{x,[t_1,t_2]} = \frac{cmP^o_{x+k,t_2}/cmP^o_{x,t_1}}{1 - (cmD^o_{x,[t_1,t_2]}/cmP^o_{x,t_1})} \tag{4.1}$$

[4]The assumption is more restrictive than we made it sound: if population coverage is not complete in either source, then the subpopulations missed in each census must be random relative to their true and reported age.

where cmP^o_{x,t_1} and $cmP^o{}_{x,t_2}$ are cumulative populations over ages x and $x+k$ in the first and second census, respectively, and $cmD^o_{x,[t_1,t_2]}$ is the cumulative deaths after age x during the intercensal period. This expression is a simple contrast between two different estimates of the same underlying quantity (population parameter), namely, the cumulative survival ratio: the denominator uses the complement of the observed ratio of (cumulative) intercensal deaths to (cumulative) population in the first census, whereas the numerator expresses it as the survival ratio computed from the cumulative counts in two successive population censuses. It is useful to express (4.1) in a logarithmic form, namely,

$$\ln(cmR_{x,[t_1,t_2]}) = \ln(SN^o_{x,x+k}) - \ln(SD^o_{x,x+k}) \tag{4.2}$$

where $SN^o_{x,x+k}$ is the 'survival ratio' computed from two censuses and $SD^o_{x,x+k}$ is the survival ratio computed from intercensal deaths[5]. In the absence of migration, age misstatement and imperfect completeness of census and death counts, both estimators should yield the same number, the ratio in (4.1) should be 1, and the log expression in (4.2) should be 0 for all adult ages.

To shed light on the meaning of expressions (4.1) or (4.2) and to simplify notation and terminology we will speak of *net age misreporting* to refer to the net result of both age over and under statement. Furthermore, because we, as well as past research, uncover systematic net age overstatement of adult ages in LAC countries, we will speak of 'age overstatement' or 'age overreporting' even though we refer to the net result of age under and over reporting. In Appendix C we show that when the assumption of absence of age misreporting is violated, we can approximate (4.2) as

$$\ln(cmR_{x,[t_1,t_2]}) \sim \ln\left(\frac{h(x+k)}{h(x)}\right) - \left(\frac{g(x)}{h(x)} - 1\right)\left(1 + I^T_{x,x+k}\right) \tag{4.3}$$

where $I^T_{x,x+k}$ is a true integrated hazard analogue between ages x and $x+k$ (and hence strictly positive), $h(x)$ is an increasing function of age that depends on age overstatement of populations and $g(x)$ is an increasing function of age that depends only on overstatement of ages at death. Both $h(x)$ and $g(x)$ are functions of the propensity to overstate and the underlying population and deaths age distribution. Assume now that the propensity to overstate ages (of populations or deaths) is age invariant or increases with age and that the following three conditions hold: (a) the (true) age distribution slopes sharply downward, (b) the age distribution of deaths increases with age, and (c) the rate of decrease of population with age is smaller that the rate of increase of deaths with age. Under these three conditions, almost universally verified in all human populations, the ratio $h(x+k)/h(x)$ will always be larger than 1 and will increase with age, $g(x)$ will always be larger than 1 and increase with age, and the rate of increase in $g(x)$ will exceed the rate of increase in $h(x)$ so that $g(x) > h(x)$ almost everywhere in the age span. The following are possible scenarios[6]:

1. When there is systematic age overstatement of population counts ONLY, $h(x) > 1$ and $g(x) = 1$, then expression (4.3) reduces to

$$\ln(cmR_{x,[t_1,t_2]}) = \ln\left(\frac{h(x+k)}{h(x)}\right) + (h^{-1}(x) - 1)(1 + I^T_{x,x+k}) < 0$$

[5]In Appendix C we provide terminology and a full justification for the use of this index.

[6]The impact of age misreporting predicted analytically in these scenarios has been confirmed by simulation studies (Condran et al., 1991; Palloni and Pinto, 2004; Grushka, 1996). In section 6 we show that our simulations also accord with analytic predictions.

The inequality results because the positive term in the expression, that is, the distortion of the survival ratio based on population counts, will be smaller than the negative term influenced by the distortion in the second estimator based on intercensal death rates.

2. When there is systematic age overstatement of death counts ONLY, $h(x) = 1$ and $g(x) > 1$, the expression becomes

$$\ln(cmR_{x,[t_1,t_2]}) = \ln\left(\frac{h(x+k)}{h(x)}\right) + (g(x) - 1)(1 + I^T_{x,x+k}) > 0$$

and the positive sign results from the fact that all terms in the expression are positive.

3. When there is systematic overstatement of BOTH population and death counts, $g(x) > h(x) > 1$, then

$$\ln(cmR_{x,[t_1,t_2]}) = \ln\left(\frac{h(x+k)}{h(x)}\right) + \left(\frac{g(x)}{h(x)} - 1\right)(1 + I^T_{x,x+k}) > 0$$

because, by assumption, all terms are positive.

Before we can use the above to diagnose conditions in an empirical case, two issues must be resolved. First, it is possible that there are empirical patterns of age overstatement of deaths and populations that offset each other and produce ratios close to 1 even though the underlying data are incorrect. That is, scenario (3) is such that the log of the ratio is 0 at all ages even when there is net age overstatement. Because of this possibility, a diagnostic of observed conditions based on the index (or the log of the index) can only detect consistency (including error consistency) of age declaration in population and death counts, rather than suggest accuracy (Dechter and Preston, 1991). Second, throughout we assumed that both census and death counts had perfect coverage. When one allows for defective census coverage, an identification problem is created since now we will have

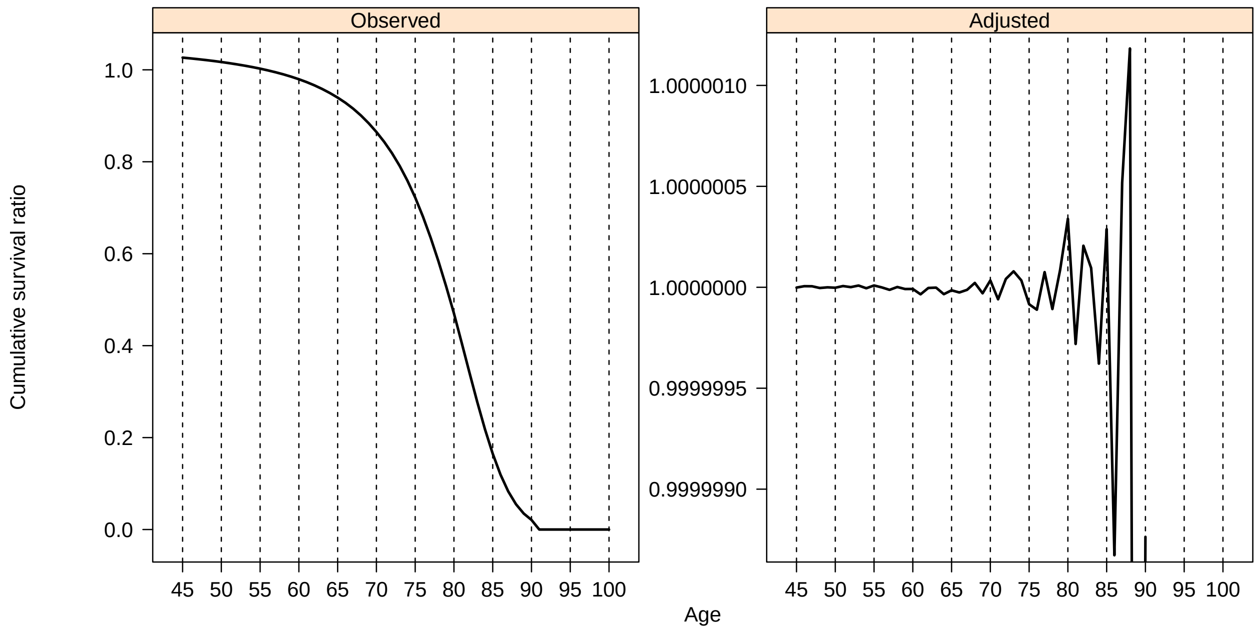
$$\ln(cmR_{x,[t_1,t_2]}) \sim \ln\left(\frac{C_2}{C_1}\right) + \ln\left(\frac{f(x+k)}{f(x)}\right) - \left(\frac{C_3 \cdot g(x)}{C_1 \cdot h(x)} - 1\right)(1 + I^T_{x,x+k}) \tag{4.4}$$

and it is clear that we can no longer separate the role of age overstatement and completeness. In particular, even if there is no age misreporting, expression (4.4) can yield non-zero values and mimic increasing or decreasing patterns with age that result naturally from age overstatement alone. To understand better the combined influence of defective coverage and age misreporting on observed mortality rates we need to define more precisely the nature of the functions $h(x)$ and $g(x)$, the nature of their dependence on patterns of age misreporting and how they interact with defective coverage. We investigate this issue in the section below.

4.2 Correction of errors due to age misreporting

As indicated before, the main tool to detect adult age misreporting is highly sensitive to relative completeness of census counts. Figure 4 displays the value of cmR_x that one obtains when there is no age misreporting at all but there is differential completeness in census counts. Thus, one cannot learn much about patterns of age misreporting unless population census counts are first adjusted. This requires to identify methods that provide robust estimates of completeness of one census relative to the other. As we show below, the evaluation study confirms a result first noted

Figure 4: Behavior of index of age misstatement with differential censuses.



by Ken Hill (Hill et al., 2009) and shows that the modified Brass technique (Brass-Hill) produces a robust estimate of C_1/C_2. The ratio of completeness factor is sufficient to correct the observed values of cmR_x.

Once the ratios are adjusted there remains the task of retrieving estimates of the magnitude of net adult age net overstatement. The model developed before based on a known standard of age net overreporting includes two parameters, λ^{no} and ϕ^{no} for the magnitude of population age over and understatement, respectively. There are three different methods to estimate these parameters.

i *A brute force method*: it is possible, but not advisable or even necessary (see (ii) below), to use the cumbersome but exact procedure that consists of computing the values for the vector $[cmR_{x=45,100}]$ that can be generated by combinations of the known vectors $[\alpha_{1x=45,100}]$ and $[\alpha_{2x=45,100}]$ and multiple pairs $(\lambda^{no}, \phi^{no})$ and then choose the (unique) pair of values that best reproduces the observed vector $[cmR_{x=45,100}]$

ii *Parametric method I*: this method is a short cut for Method I. We used simulated data to estimate the following relation

$$(cmR_x)^{-1} = \alpha_{0x} + \alpha_{1x}\lambda^{no} + \alpha_{3x}\phi^{no} \qquad (4.5)$$

for all values of $x \geq 45$. The parameters of this relation, α_0, α_1 and α_3, characterize the space of solutions for the triplet $(cmR_x, \lambda^{no}, \phi^{no})$ embedded in the simulated data. As shown below in Table 5 the fit of the model is very good and the estimated values of the constant is always close to 1, as it should be. If the observed data is an element of the space of solutions, that is, if the observed data is generated by one of the combinations of parameters that spawns the simulation, it might be possible to invert the procedure in (4.5), use the coefficients estimated from (4.5) and compute the pair of values (λ^{no}, ϕ^{no}) that reproduces the observed value $cmR_x - 1$ for all x[7]. We show in Table 6 that given an observed vector of values $\{cmR_{x=45,100}\}$ and the vectors of parameters $\{\alpha_{1x=45,100}\}$ and $\{\alpha_{3x=45,100}\}$ there is a unique and best (in mean squared error sense) solution for the unknown parameters of model (4.5) [8].

iii *Parametric method II*: the third method seeks to reproduce the shape of the function $[cmR_{x=45,100}]$ as a function of age and then map parameters of the function onto the pairs (λ^{no}, ϕ^{no}) that generated the data. It consists of fitting a hyperbola to a range of values of cmR_x

$$cmR_x = \beta_1/(\varsigma - age)^{\beta_2} \tag{4.6}$$

where ς is set equal to 76[9]. We then use the estimated parameters of function (4.6) to predict the pair of values $(\lambda^{no}, \phi^{no})$. As we show in Table 7 the fit of the hyperbolic function to the distorted data is very tight but the retrieval of the hidden parameters governing net age overstatement is generally poor. This is due to under-identification: if one uses the entire range of values attainable by λ^{no} and ϕ^{no}, the function $cmR_{x=45,100}$ can be mapped onto multiple pairs (λ^{no}, ϕ^{no}). The procedure works best when the pair of values (λ^{no}, ϕ^{no}) is within a limited range (approximately [0.10-1.5]). Because of this regularity one can use method (ii) and (iii) jointly to seek consistency: if the observed values of the parameters λ^{no} and ϕ^{no} are within the identification range, then both methods should produce the same results.

5 Evaluation study

The nature of problems generated by faulty national vital statistics and censuses is highly heterogeneous and vary by country, time period, age groups, gender and surely by regions. This is complicated by the fact that there are multiple techniques or procedures, each relying on specialized assumptions, to adjust for errors that exist in the data. Over the last two to three decades, but mostly in the late seventies and eighties, demographers developed a large number of techniques to adjust faulty data from censuses, vital statistics and population surveys to estimate both fertility and mortality. There are nearly 15 different, albeit not completely independent methods, to correct for completeness errors (but not age misreporting) of adult mortality statistics, each with its own peculiar advantages and shortcomings, and each depending on sets of different but overlapping assumptions.

Optimal adjustments for faulty coverage and age misreporting are unfeasible in the absence of well-established criteria to decide which candidate techniques performs optimally and under which conditions they do or do not do so. To assess the performance of alternative procedures and

[7] The constrain imposed, namely, that the observed data must be in the space of populations generated by the simulation is crucial for in the simulation we do not use all possible values of $(\lambda^{no}, \phi^{no})$ but we limit them to a rather small range.

[8] Model (4.5) is best fitting in the sense that any interaction terms or higher order moments of the independent variables do not reduce the mean squared error by a statistically significant amount.

[9] In cases when the values of the magnitude of age overstatement approaches the largest values allowed (close to 2 or 2.5), the function cmR_x attains a point of discontinuity where the derivatives with respect to age do not exist. In order to avoid such cases we used trial values for the parameter ς and find that, in the space of simulated populations, $\varsigma = 76$ is optimal as it always avoids points of discontinuity. This is equivalent to saying that one cannot reproduce the function for ages above 76, a trait that is partially responsible for under identification.

Table 5: Regression model relating index of age misstatement and parameters of age misreporting.

Age	α_0	α_1	α_2	R^2
45	1.000	-0.027	-0.004	1.000
46	1.000	-0.012	-0.005	1.000
47	1.000	-0.006	-0.005	1.000
48	1.000	-0.003	-0.006	1.000
49	1.000	0.000	-0.007	1.000
50	1.000	0.002	-0.008	1.000
51	1.000	0.003	-0.009	1.000
52	1.000	0.005	-0.010	1.000
53	1.000	0.006	-0.011	1.000
54	1.000	0.008	-0.013	1.000
55	1.000	0.010	-0.014	1.000
56	1.000	0.012	-0.016	0.999
57	0.999	0.014	-0.019	0.999
58	0.999	0.017	-0.022	0.999
59	0.999	0.020	-0.025	0.999
60	0.999	0.024	-0.030	0.999
61	0.999	0.029	-0.035	0.999
62	0.999	0.035	-0.041	0.999
63	0.998	0.042	-0.048	0.999
64	0.998	0.051	-0.057	0.998
65	0.997	0.062	-0.069	0.998
66	0.996	0.076	-0.082	0.998
67	0.995	0.094	-0.099	0.997
68	0.994	0.116	-0.121	0.997
69	0.992	0.145	-0.148	0.996
70	0.990	0.183	-0.183	0.995
71	0.986	0.231	-0.228	0.995
72	0.982	0.295	-0.285	0.994
73	0.975	0.378	-0.360	0.992
74	0.966	0.490	-0.458	0.991
75	0.952	0.638	-0.586	0.989

Table 6: Results from inverse method of age misstatement to recover parameters of age misreporting.

run	ϕ^{no}	$\hat{\phi}^{no}$	λ^{no}	$\hat{\lambda}^{no}$	R^2
1	0.000	0.061	0.350	0.370	1.000
2	0.000	0.002	0.700	0.685	1.000
3	0.000	-0.059	1.050	0.999	1.000
4	0.000	-0.118	1.400	1.313	1.000
5	0.000	-0.178	1.750	1.628	1.000
6	0.000	-0.238	2.100	1.942	1.000
7	0.000	-0.298	2.450	2.256	1.000
8	0.000	-0.358	2.800	2.571	1.000
9	0.350	0.393	0.700	0.727	1.000
10	0.350	0.392	1.050	1.078	1.000
11	0.350	0.391	1.400	1.429	1.000
12	0.350	0.390	1.750	1.780	1.000
13	0.350	0.388	2.100	2.130	1.000
14	0.350	0.387	2.450	2.481	1.000
15	0.350	0.386	2.800	2.832	1.000
16	0.700	0.710	1.050	1.067	1.000
17	0.700	0.755	1.400	1.445	1.000
18	0.700	0.801	1.750	1.823	1.000
19	0.700	0.846	2.100	2.201	1.000
20	0.700	0.892	2.450	2.579	1.000
21	0.700	0.938	2.800	2.957	1.000
22	1.050	1.013	1.400	1.393	1.000
23	1.050	1.096	1.750	1.791	1.000
24	1.050	1.179	2.100	2.189	1.000
25	1.050	1.262	2.450	2.587	1.000
26	1.050	1.345	2.800	2.985	1.000
27	1.400	1.303	1.750	1.704	1.000
28	1.400	1.416	2.100	2.117	1.000
29	1.400	1.530	2.450	2.530	1.000
30	1.400	1.643	2.800	2.943	1.000
31	1.750	1.582	2.100	2.004	0.999
32	1.750	1.720	2.450	2.427	1.000
33	1.750	1.859	2.800	2.851	1.000
34	2.100	1.851	2.450	2.292	0.999
35	2.100	2.009	2.800	2.723	1.000
36	2.450	2.110	2.800	2.569	0.998

Table 7: Non-linear regression to recover parameters of age misreporting.

run	ϕ^{no}	$\hat{\phi}^{no}$	λ^{no}	$\hat{\lambda}^{no}$	R^2
1	0.000	1.243	0.350	0.071	1.000
2	0.000	1.668	0.700	0.171	0.998
3	0.000	2.662	1.050	0.339	0.988
4	0.000	8.583	1.400	0.890	0.936
5	0.000	11.785	1.750	0.952	0.918
6	0.000	9.955	2.100	0.819	0.937
7	0.000	20.000	2.450	4.240	0.925
8	0.000	26.352	2.800	1.186	0.905
9	0.350	1.244	0.700	0.070	1.000
10	0.350	1.627	1.050	0.160	0.998
11	0.350	2.428	1.400	0.303	0.991
12	0.350	5.470	1.750	0.639	0.958
13	0.350	7.273	2.100	0.721	0.946
14	0.350	24.000	2.450	5.584	0.986
15	0.350	43.669	2.800	1.533	0.886
16	0.700	1.245	1.050	0.069	1.000
17	0.700	1.593	1.400	0.152	0.998
18	0.700	2.264	1.750	0.275	0.994
19	0.700	4.228	2.100	0.519	0.973
20	0.700	73.344	2.450	3.738	0.993
21	0.700	45.485	2.800	1.833	0.906
22	1.050	1.245	1.400	0.068	1.000
23	1.050	1.565	1.750	0.144	0.999
24	1.050	2.142	2.100	0.253	0.995
25	1.050	3.562	2.450	0.445	0.981
26	1.050	13.985	2.800	1.235	0.939
27	1.400	1.246	1.750	0.067	1.000
28	1.400	1.542	2.100	0.138	0.999
29	1.400	2.047	2.450	0.236	0.996
30	1.400	3.149	2.800	0.394	0.986
31	1.750	1.246	2.100	0.066	1.000
32	1.750	1.522	2.450	0.132	0.999
33	1.750	1.972	2.800	0.221	0.997
34	2.100	1.246	2.450	0.065	1.000
35	2.100	1.504	2.800	0.127	0.999
36	2.450	1.246	2.800	0.064	1.000

choose an optimal adjustment strategy we develop an evaluation study designed to identify best adjustments for relative completeness and age misreporting. The goal of the study is to generate distributions of errors associated with each adjustment procedure under a diverse set of conditions that violate the assumptions of which the procedures rely[10]. Thus, not only can we choose the optimal adjustment technique under a given set of (observed) conditions, e.g. the one minimizing some error functions, but we can also assess the magnitude of errors when a combination of these assumptions is violated. Our evaluation study is similar to and extends the work of Hill and colleagues (Hill et al., 2009; Hill and Choi, 2004; Hill, 2003; Hill et al., 2005).Our study includes 11 methods, considers adult age misreporting[11], and produces distributions of errors associated with each adjustment technique when (known) combinations of assumptions are violated.

The evaluation study proceeds as follows: we first simulate populations representing different demographic profiles (stable, quasi-stable and non-stable) driven by combinations of (a) constant fertility and mortality, (b) constant fertility and declining mortality, and (c) declining fertility and declining mortality. We then combine these profiles with different patterns of distortions due to faulty coverage of population and death counts and adult age misreporting. A battery of 11 techniques is deployed and in each case we compute multiple measures of performance comparing the true parameter(s) with those retrieved by each technique. We rank the performance of techniques for each combination of conditions violating assumptions on which the techniques rely. Finally, we score techniques according to their sensitivity to violation of combinations of assumptions. The optimal technique is then paired with a new procedure to adjust for age misreporting and, jointly, they are used in an algorithm to make final adjustments to observed adult mortality rates. A crucial issue discussed below is the order in which these techniques, one for adjustment of coverage and one for age misreporting, must be deployed and the justification for that order.

5.1 Simulated populations: five classes of demographic profiles

We first simulate a large number of populations spanning a broad range of fertility and mortality regimes that come close to reproducing age-specific counts of deaths and populations that would have been observed over an interval of about 100 years in the absence of errors in the data. We start out with a stable age distribution in single years of age, e.g., $P_{xt_0}, x = 0, ...100$, to represent an average population in 1900 and then project it forward for 100 years using schedules of mortality, e.g. $(S_x = 0, 100)$, and fertility, e.g., $(F_x = 15, 50)$ [12]. We chose four different trajectories of mortality and fertility roughly reproducing four classes of demographic transitions experienced by Argentina, Costa Rica, Guatemala and Mexico respectively (Palloni, 1990). All four trajectories are defined by choosing values of life expectancy at birth (E_0), and Gross Reproduction Rate (GRR) thus identifying the rate of natural increase (r) for every decade between 1900 and 2000. With the exception of the first trajectory (corresponding to the experiences of Argentina and Uruguay), we assume an initial stable populations with r and E_0 equal to those observed in the first population census before 1940 for each trajectory. In the case of the Argentina/Uruguay profile we use the observed average age distribution in the population censuses within the period 1850-1910. We assume linear intra-decade changes in the two key population parameters, r and

[10] The investigations that follow were first documented elsewhere (Palloni and Pinto, 2000)

[11] Hill and colleagues did consider simulations that included limited forms of age misreporting. We augment this aspect to capture patterns of age misreporting typically observed in LAC countries as well as the performance of a new method to adjust for associated errors.

[12] Throughout we use conventional mathematical notation and when referring to discrete functions we employ subscripts, e.g. P_x, whereas for continuous functions we use the parentheses enclosing the function's argument, e.g. *P(x)*.

E_0 and, additionally, that each type of demographic transition profile preserves the age patterns of mortality and fertility. We chose the West model in the Coale-Demeny family of life tables and an age pattern of fertility identical to the one used in the computations of the Coale-Demeny stable population models (Coale et al., 1983). Information on the four classes of demographic transitions used here are in Appendix A. Finally, we construct a fifth profile of a stable population with natural rate of increase and fertility pattern equivalent to the average of LAC populations in the interval 1950-60, e.g. not yet heavily perturbed by large scale net migration as is the case in Argentina, Brazil, Cuba, and Uruguay, or early fertility changes, as in Argentina and Uruguay.

Following routine population projection calculations we produce 505 populations and associated distributions of births and deaths by single calendar year and single years of age. The simulated populations represent a very broad set of experiences, from those preserving population stability up until 1950 or thereabouts, to those shifting to quasi-stability from 1930 up to 1980, to those with little or no stability at all from the start[13].

5.2 Simulated distortions I: imperfect relative completeness of death registration

Distortions due to population or death coverage can be implemented in a straightforward matter. We define observed population (or death) counts by age as a fraction of the simulated (true) quantities:

$$\begin{aligned} P^o_{xt_1} &= C_1 P^s_{xt_1} \\ P^o_{xt_2} &= C_2 P^s_{xt_2};\ t_2 < t_1 \\ D^o_{xt} &= C_3 D^s_{xt};\ t = t_1, t_1+1, ... \leq t_2 \end{aligned}$$

for $x \geq 5$, where $P^o_{xt_1}$ is the observed (distorted) population at age $(x, x+1]$ at time t_1, $P^o_{xt_2}$ is the observed (distorted) population at age $(x, x+1]$ at time t_2, and D^o_{xt} is the observed (distorted) number of deaths in year t; $P^s_{xt_1}, P^s_{xt_2}$ and D^s_{xt} are the simulated (true) quantities and C_1, C_2 and C_3 are the fractions of total events actually observed (completeness factors). The completeness factors for censuses were set at values in the range 0.80-1.0 in intervals of 0.5 whereas the death completeness factors varied between 0.70 and 1.0 in intervals of 0.5. Altogether we produce a total of 875 (175*5) patterns of including distorted and true demographic profiles. These definitions are sufficient to evaluate adjustment methods that require only one census and one to three years of death counts centered on the census or, alternatively, those that demand as inputs two population censuses and an array of intercensal deaths.

The above set up contains a massive assumption, namely, that completeness of both population and death counts is age invariant. At least within the age range in which the techniques are deployed (5-85), the assumption is unlikely to be met, particularly for population counts. To complete the set of reasonable distortions we add two different patterns of age varying completeness generating a total of 2,625 simulated populations. We show later, however, that as long as the difference between maximum and minimum completeness stays below 10% of the mean value of completeness, the variance of completeness by age does not have a strong impact on choices of techniques (Section 6).

[13]To compute single years of age stable populations we first generate single years of age life tables by respecting the separator factors adopted by Coale and Demeny and the use of standard stable population expressions. The precise routine followed is in a STATA do file available on request from authors.

5.3 Simulated distortions III: combining age misreporting and faulty coverage

We now have all the ingredients to generate distorted populations using as benchmarks the demographic profiles described above. The defective populations were defined considering each demographic profile separately, letting C_1 and C_2 take on values between 0.80 and 1.0 in intervals of 0.05 whereas C_3 takes on values between 0.75 and 1.0 in intervals of 0.05 and, finally, assigning values to ϕ^{no} and λ^{no} ranging from 0 to 2.5 in intervals of 0.50. We use all possible combinations of these parameters and generate a total of 6,300 populations per demographic profile (5 in all) for a population space containing a total of 31,500 observations or populations in single years of age traced for a total of 100 years. In addition, to test for sensitivity to violations of the assumption of age invariant relative completeness, we add two patterns of deviations and generate a space of 94,500 populations.

5.4 Application of adjustment techniques

The next stage in the evaluation is to apply the 11 techniques to adjust for defective completeness as well as the technique developed above to correct for age misreporting.

5.4.1 Techniques to adjust for defective completeness

The most important techniques to detect and adjust for faulty completeness evaluated in this study are summarized in Table 8[14]. The table identifies techniques using the names of researcher(s) who proposed them or modified an original version. The table highlights (a) key assumptions on which the techniques rely, and (b) information required to implement each of them. These methods share important commonalities and all but two (Brass No 1 and Preston-Hill No 1) abstain from invoking the assumption of stability. Yet they differ in at least one feature that, under suitable empirical conditions, grants them an advantage over competing methods.

The key features of these techniques are the following:

- Computation of rates of growth: with two exceptions (Preston-Hill No1 and Brass) all methods require computation of age specific rates of growth in an intercensal period. Because observed rates may be perturbed by differential census completeness, the estimates of the main parameter (relative completeness of death registration) could be biased if the method is sensitive to differential census completeness. A way around this is to first adjust for relative completeness of census registration and then apply any of the techniques using adjusted age specific rates of growth. This idea was first put forward by Hill (Hill and Choi, 2004; Hill et al., 2009) who suggests that one of the methods listed in the table (Brass-Hill) be used to retrieve a robust estimate of the ratio of completeness of both censuses.

- Population closed to migration: none of the methods in Table 8 works well in the presence of significant intercensal migration. If information on net migration is available, it must be used to adjust the observed rates of intercensal growth[15]

- Absence of age misreporting: all methods assume either no age misreporting or, alternatively, age misreporting that perturbs only trivially the figures of cumulative population above adult

[14] We reviewed a longer list of techniques and, with two exceptions, chose to test only those that did not rely on the assumption of stability or quasi-stability.

[15] Hill and colleagues investigated the effects of intercensal migration (Hill et al., 2009). In the simulations performed here we do not include consideration of migration but its effects are partially captured via differential censuses completeness.

Table 8: Methods to adjust for completeness of death registration: assumptions and required data.

Method	Assumptions	Required Data
Brass (B)	1-2-3-4-5	B
Brass-Hill (BHill[2])	2-3-4	A
Brass-Martin (BMartin[3])	1-2-3-4-6	B
Bennet-Horiuchi No 1 (BH_1)	1-2-3-4	A
Bennet-Horiuchi No 2 (BH_2)	1-2-3-4	A
Bennet-Horicuhi No 3 (BH_3)	1-2-3-4	A
Bennet-Horiuchi No 4 (BH_4)	1-2-3-4	A
Bennet-Horiuchi No 5 (2SBH_4)	1-2-3-4	A
Preston-Hill No 1 (PH_1)	1-2-3-4-5	B
Preston-Hill No 2 (PH_2)	1-2-3-4	A
Preston-Bennet (PB)	1-2-3-4	A
Preston-Lahiri No 1 (PL_1)	1-2-3-4	A
Preston-Lahiri No 2 (PL_2)	1-2-3-4	A

[1]See appendix 5 for definitions of the four variants of Bennet-Horiuchi method and the two variants of Preston-Lahiri method.
[2]BHill is a method we use to retrieve estimates of the ratio of completeness of the first relative to the second census.
[3]BMartin is a variant of Brass classic method that relaxes the assumption of stability and assumes instead past mortality decline.

KEYS FOR ASSUMPTIONS
1. Identical completeness of census counts in both census
2. Closed to migration
3. No age misreporting
4. Invariant completeness by age
5. Stability
6. Quasi stability

KEYS FOR REQUIRED DATA
A. Two censuses and intercensal deaths
B. One census and one to three years of deaths by age

ages. This poses a conundrum: if, as asserted before, LAC population and mortality counts are heavily affected by age overstatement, how can one expect to obtain precise estimates of relative completeness using techniques that are vulnerable when there is age misreporting? There are two conditions that provide a escape from this trap. The first is that the type of age misreporting that predominates in LAC is net age overstatement. When using cumulative populations over some age x the damage done to the target quantity by age misreporting only depends on population flows across age x originating at younger ages. It is insensitive to transfers of population above age x. Furthermore, the relative volume of flows, e.g. the relative error of the target quantity, is generally low for late adulthood and early old ages (less than 65 or 70) though it begins to mount after age 75 or so. Since in all cases computations only require to employ observations up to ages 70 or 75, the impact of age overstatement will be minor[16].The second favorable condition that circumvents the problem is that the optimal method (Bennett-Horiuchi No 4) is also the least sensitive to age misreporting of the type encountered in LAC (see below).

- Age invariant relative completeness of death registration: all techniques rely on the assumption that the relative completeness of death registration is age invariant. However, as we show later, when there are mild violations of the assumption the optimal method we choose (Bennett-Horiuchi IV) performs best.

- Estimation of life expectancy at older ages: all methods adopt *ad hoc* procedures to handle the open age group. These procedures rely on exogenous computations of parameters relating the quantity of interest, life expectancy at age 75 or 70 and selected observed quantities in the data at hand. The relations are estimated using model life tables, stable population expressions, numerical approximations or a combinations of all these. In the applications implemented here we follow the methods suggested by the authors in each case. Thus, some of the variability in performance that we uncover, albeit a small part, is due to heterogeneous strategies to handle the open age group.

5.4.2 Techniques to adjust for age misreporting

We consider only one technique to adjust the observed data for age misreporting. As described before, the procedure rests on two key assumptions. The first is that errors follow a known age pattern (the Costa Rican standard). The second is the age pattern of age misreporting is the same in the census and in vital statistics. both are simplification and a more comprehensive evaluation study should include deviant patterns.

6 Results of the evaluation study

We now review results of applying candidate techniques for adjusting defective relative completeness and age misreporting. We base our discussion on results from the set of simulated populations describe before, a space of fictitious populations and deaths generated by five different demographic regimes combined with an exhaustive set of error patterns. In section 6.1 we describe the behavior of these techniques, that is, their effectiveness to retrieve population parameters under several conditions: ignoring the error patterns embedded in the space of simulated populations, in subsets

[16]This is because even with heavy age overstatement the population at any particular age $y < x$, where x is below 65 or so, is a small fraction of the population above age x. These ratios increase as x increases due to exponential decrease of population at older ages.

of populations defined by selected underlying conditions and, finally, isolating two types of errors that violate basic assumptions of all methods considered here, namely, age misreporting and age dependent completeness. In section 6.2 we describe the behavior of methods to adjust for age misreporting.

6.1 Defective completeness: evaluation using pooled simulated populations

To facilitate assessment of techniques we create six different populations subsets: (a) total or pooled, (b) stable, (c) non-stable, (d) non-stable with no age misreporting, with defective death and population coverage, (e) non-stable with age misreporting, incomplete death coverage and defective but identical population coverage in the two censuses and (f) non-stable with age misreporting, incomplete death and population coverage. Each subpopulation with incomplete population and/or death coverage has three variants, one with constant relative completeness (of census and deaths counts) and the others with age varying completeness.

Investigating the behavior of techniques isolating conditions that generate errors is helpful when there is reliable external information about population stability, nature of age misreporting and/or patterns of age relative completeness. A technique that performs optimally in the pooled simulated population may not do so well under a specific set of conditions. The opposite situation is also possible: a technique may not behave well on average but could be optimal under some circumstances. Because the source of uncertainty matters for the final choice of method, our assessment is carried out across multiple subsets of simulated populations, each reflecting different types of errors or conditions. We define the following six population subsets: a) pooled sample (n=31,500), b) stable populations (n=6,300), c) non-stable populations (n=25,200), d) non-stable populations with no age misreporting but defective completeness of death and population counts (n=700) e) non-stable populations with age misreporting, defective coverage of death counts and equal (possible defective) coverage of population counts (n=4,320) and, finally, f) non-stable population with age misreporting, defective death registration, defective (but unequal) population counts (n=17,280). In each of these subsets we generate three variants, one assuming constant relative completeness and two variants imposing two different age-dependent patterns of relative completeness[17].

We evaluate the following techniques: Brass technique (Brass, 1975) modified by Hill (Hill, 1987) to compute a robust estimate of relative completeness in two population censuses and 11 techniques to estimate relative completeness of death registration: a) original Brass method (Brass, 1975) modified by Hill (Hill, 1987) and variant by Martin (Martin, 1980), b) four variants of Bennett and Horiuchi (Bennett and Horiuchi, 1981;1984), c) one method by Preston and Bennett (Preston and Bennett, 1983), d) two different methods by Preston and Hill (Preston and Hill, 1980), and e) two variants of Preston and Lahiri (Preston and Lahiri, 1991).

The assessment focuses on the mean proportionate (absolute) errors for two population parameters, the ratio of completeness of first to second census coverage, ρ^c=C_1/C_2 and the relative completeness of death registration, $\rho^d = (C_3/(.5*(C_1+C_2))$. Tables 9–11, panels A through panel F display the mean of the proportionate absolute error for each of the six populations subsets defined above. The errors in each population subset $s, s = 1, 2...6$, are $\Xi_s^d = \sum_{j=1}^{j=K_s} \varepsilon_{sj}^d$ and $\Xi_s^c = \sum_{j=1}^{j=K_s} \varepsilon_{sj}^c$ where $\varepsilon_{sj}^d = \mid \hat{\rho}_{sj}^d - \rho_{sj}^d \mid / \rho_{sj}^d$, $\varepsilon_{sj}^c = \mid \hat{\rho}_{sj}^c - \rho_{sj}^c \mid / \rho_{sj}$, ρ^c and ρ^d are defined as before, $\hat{\rho}_{sj}^d$ and $\hat{\rho}_{sj}^c$ are estimates, and the summations are over all simulated populations j in each of six

[17] The two functions for age dependent census completeness are assumed to hold in both censuses and are defined as follows: **(a) scenario 1**: C_1= 0.75 if age [15-34] and C_1= 0.85 elsewhere; C_2=0.85 if age [15-34] and C_2= 0.95 elsewhere; C_3= 0.80 if age [15-34] and C_3= 0.85 elsewhere; **(b) scenario 2**: C_1= 0.85 if age [15-34] and C_1= 0.75 elsewhere; C_2= 0.95 if age [15-34] and C_2= 0.85 elsewhere; C_3= 0.85 if age [15-34] and C_3= 0.80 elsewhere.

subsets. Naturally, different error metrics yield different ranking of methods but the measure we use is the preferred one in most applications of this kind.[18]

The six panels of Tables 9–11 display the mean of the proportionate absolute error for each of the six populations subsets defined above. Table 9 refers to simulations with constant relative completeness by age and Tables 10 and 11 reflect results using two different patterns of age varying relative relative completeness. The errors in each population subset $s, s = 1, 2...6$, are $\Xi_s^d = \sum_{j=1}^{j=K_s} \varepsilon_{sj}^d$ and $\Xi_s^c = \sum_{j=1}^{j=K_s} \varepsilon_{sj}^c$, where $\varepsilon_{sj}^d =\mid \hat{\rho}_{sj}^d - \rho_{sj}^d \mid / \rho_{sj}^d$, $\varepsilon_{sj}^c =\mid \hat{\rho}_{sj}^c - \rho_{sj}^c \mid / \rho_{sj}$, ρ^c and ρ^d are defined as before, $\hat{\rho}_{sj}^d$ and $\hat{\rho}_{sj}^c$ are estimates, and the summations are over all simulated populations j in each of six subsets. Naturally, different error metrics yield different ranking of methods but the measure we use is the preferred one in most applications of this kind[19].

[18]The figures in Tables 9–11, panel A through panel F are computed using a subset of rather benign patterns of distortions as they exclude values of completeness lower than 0.7 and differences between completeness of successive censuses higher than 0.10.

[19]We emphasize that the figures in Tables 9–11, are computed on a subset of rather benign patterns of distortions as they exclude values of relative completeness lower than 0.7 and differences between completeness of successive censuses higher than 0.10.

Table 9: Proportionate absolute errors in each of six populations subsets with age invariant relative completeness.

	A. Stable & Nonstabe			B. Stable			C. Nonstable			D. Nonstable*			E. Nonstable•			F. Nonstable‡		
Indicator	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
Brass Hill Census (BHill)[1]	0.003	0.003	0.003	0.005	0.005	0.004	0.003	0.003	0.002	0.001	0.001	0.001	0.002	0.003	0.002	0.002	0.003	0.002
Bennet Horiuchi No 1 (BH_1)	0.242	0.304	0.265	0.199	0.263	0.224	0.251	0.314	0.273	0.215	0.294	0.251	0.010	0.014	0.011	0.212	0.256	0.117
Bennet-Horiuchi No 2 (BH_2)	0.248	0.300	0.256	0.215	0.260	0.216	0.260	0.310	0.264	0.215	0.296	0.253	0.011	0.013	0.010	0.219	0.256	0.111
Bennet-Horicuhi No 3 (BH_3)	0.240	0.303	0.263	0.200	0.264	0.225	0.247	0.312	0.271	0.212	0.293	0.249	0.010	0.013	0.011	0.210	0.255	0.115
Bennet-Horiuchi No 4 (BH_4)	0.248	0.300	0.256	0.215	0.260	0.216	0.260	0.310	0.264	0.215	0.296	0.253	0.011	0.013	0.010	0.219	0.256	0.111
Bennet-Horiuchi No 5 (2SBH_4)	0.021	0.024	0.017	0.016	0.020	0.015	0.023	0.025	0.017	0.007	0.008	0.005	0.022	0.024	0.016	0.023	0.025	0.017
Brass-Martin(BMartin)[2]	0.079	0.107	0.085	0.038	0.038	0.021	0.110	0.124	0.086	0.057	0.071	0.061	0.112	0.124	0.084	0.111	0.124	0.085
Brass Hill (BHill)[1]	0.043	0.046	0.027	0.038	0.038	0.021	0.045	0.048	0.028	0.005	0.006	0.004	0.045	0.048	0.028	0.045	0.048	0.028
Preston Bennet (PB)	0.629	0.728	0.552	0.493	0.623	0.594	0.701	0.754	0.537	0.581	0.692	0.541	0.031	0.051	0.049	0.629	0.853	0.510
Preston Hill I (PH_1)	0.340	0.388	0.297	0.275	0.381	0.375	0.356	0.390	0.274	0.358	0.388	0.267	0.203	0.226	0.146	0.325	0.308	0.175
Preston-Hill 2 (PH_2)	0.367	0.386	0.272	0.249	0.367	0.320	0.374	0.391	0.258	0.377	0.390	0.251	0.242	0.258	0.146	0.348	0.315	0.181
Preston Lahiri No 1 (PL_1)	0.406	5.911	260.880	0.336	1.498	4.478	0.449	7.014	291.655	0.452	3.434	20.699	0.021	0.023	0.015	0.423	11.192	451.144
Preston-Lahiri No 2(PL_2)	0.378	5.560	168.422	0.307	1.366	4.558	0.415	6.609	188.274	0.414	2.064	6.947	0.022	0.027	0.021	0.394	0.916	3.422
N	31,500			6,300			25,200			700			4,320			10,368		

SD, standard deviation; Med, median.

* $\theta_1 = \theta_3 = 0$

• $C_1 = C_2$ and $C_3 < 1$

‡ $C_1 \neq C_2$ and $C_3 < 1$ and maxabs$(C_1 - C_2) < .10$

[1] Values of errors in the Brass-Hill shown in the first row correspond to errors associated with the ratio C_1/C_2. While values of Brass-Hill in the seventh row correspond to erros associated with relative completeness of death registration.

[2] BMartin is a variant of Brass classic method that relaxes the assumption of stability and assumes instead past mortality decline.

Table 10: Proportionate absolute errors in each of six populations subsets with age dependent relative completeness (Scenario 1).

	A. Stable and Nonstabe			B. Stable			C. Nonstable			D. Nonstable*			E. Nonstable•			F. Nonstable‡		
Indicator	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
Brass Hill Census (BHill)[1]	0.040	0.046	0.033	0.039	0.045	0.032	0.040	0.047	0.033	0.039	0.045	0.032	0.013	0.015	0.010	0.038	0.037	0.020
Bennet Horiuchi No 1 (BH_1)	0.267	0.319	0.279	0.273	0.331	0.296	0.265	0.316	0.275	0.217	0.295	0.253	0.016	0.020	0.018	0.208	0.256	0.125
Bennet-Horiuchi No 2 (BH_2)	0.267	0.315	0.269	0.271	0.327	0.285	0.265	0.312	0.264	0.217	0.297	0.254	0.019	0.023	0.019	0.215	0.255	0.116
Bennet-Horicuhi No 3 (BH_3)	0.267	0.317	0.277	0.272	0.329	0.293	0.263	0.314	0.272	0.214	0.294	0.251	0.017	0.021	0.018	0.205	0.255	0.123
Bennet-Horiuchi No 4 (BH_4)	0.267	0.315	0.269	0.271	0.327	0.285	0.265	0.312	0.264	0.217	0.297	0.254	0.019	0.023	0.019	0.215	0.255	0.116
Bennet-Horiuchi No 5 (2SBH_4)	0.099	0.150	0.350	0.147	0.332	0.737	0.088	0.105	0.081	0.078	0.092	0.068	0.030	0.033	0.023	0.082	0.085	0.052
Brass-Martin(BMartin)[2]	0.162	0.259	0.292	0.154	0.209	0.185	0.164	0.271	0.312	0.141	0.207	0.217	0.161	0.170	0.104	0.129	0.212	0.211
Brass Hill (BHill)[1]	0.119	0.162	0.151	0.120	0.157	0.139	0.118	0.163	0.154	0.102	0.128	0.109	0.078	0.081	0.043	0.096	0.124	0.103
Preston Bennet (PB)	0.742	0.728	0.388	0.790	0.784	0.346	0.729	0.714	0.396	0.549	0.594	0.408	0.186	0.207	0.146	0.703	0.766	0.373
Preston Hill I (PH_1)	0.445	0.514	0.446	0.444	0.514	0.448	0.447	0.514	0.445	0.447	0.508	0.432	0.224	0.247	0.145	0.401	0.372	0.209
Preston-Hill 2 (PH_2)	0.456	0.505	0.415	0.459	0.504	0.417	0.432	0.506	0.415	0.450	0.500	0.401	0.260	0.277	0.145	0.431	0.368	0.216
Preston Lahiri No 1 (PL_1)	0.541	6.432	507.347	0.538	1.926	4.066	0.541	7.558	567.225	0.534	2.102	5.892	0.034	0.039	0.026	0.507	6.139	74.353
Preston-Lahiri No 2(PL_2)	0.477	6.582	256.106	0.439	1.424	2.206	0.481	7.872	286.320	0.482	2.509	11.876	0.044	0.051	0.035	0.452	1.605	3.491
N	31,500			6,300			25,200			700			4,320			10,368		

SD, standard deviation; Med, median.

* $\theta_1 = \theta_3 = 0$

• $C_1 = C_2$ and $C_3 < 1$

‡ $C_1 \neq C_2$ and $C_3 < 1$ and maxabs$(C_1 - C_2) < .10$

[1] Values of errors in the Brass-Hill shown in the first row correspond to errors associated with the ratio C_1/C_2. While values of Brass-Hill in the seventh row correspond to erros associated with relative completeness of death registration.

[2] BMartin is a variant of Brass classic method that relaxes the assumption of stability and assumes instead past mortality decline.

Scenario 1: $C_1 = 0.75$ if age [15-35], $C_1 = 0.85$ elsewhere; $C_2 = 0.85$ if age [15-35], $C_1 = 0.95$ elsewhere; $C_3 = 0.80$ if age [15-35], $C_1 = 0.85$ elsewhere.

Table 11: Proportionate absolute errors in each of six populations subsets with age dependent relative completeness (Scenario 2).

	A. Stable and Nonstabe			B. Stable			C. Nonstable			D. Nonstable*			E. Nonstable•			F. Nonstable‡		
Indicator	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
Brass Hill Census (BHill)[1]	0.041	0.045	0.031	0.042	0.046	0.031	0.041	0.045	0.031	0.043	0.046	0.031	0.029	0.030	0.010	0.033	0.036	0.028
Bennet Horiuchi No 1 (BH_1)	0.244	0.311	0.286	0.249	0.327	0.300	0.243	0.307	0.282	0.218	0.280	0.259	0.015	0.019	0.016	0.234	0.245	0.143
Bennet-Horiuchi No 2 (BH_2)	0.242	0.310	0.271	0.247	0.325	0.285	0.240	0.307	0.267	0.218	0.283	0.262	0.054	0.055	0.030	0.225	0.233	0.134
Bennet-Horicuhi No 3 (BH_3)	0.244	0.309	0.284	0.249	0.324	0.297	0.243	0.305	0.280	0.215	0.278	0.257	0.014	0.018	0.015	0.233	0.244	0.140
Bennet-Horiuchi No 4 (BH_4)	0.242	0.310	0.271	0.247	0.325	0.285	0.240	0.307	0.267	0.218	0.283	0.262	0.054	0.055	0.030	0.225	0.233	0.134
Bennet-Horiuchi No 5 (2SBH_4)	0.081	0.118	0.276	0.095	0.247	0.586	0.074	0.086	0.062	0.078	0.083	0.052	0.034	0.037	0.025	0.060	0.063	0.041
Brass-Martin(BMartin)[2]	0.114	0.154	0.155	0.109	0.124	0.089	0.116	0.162	0.167	0.105	0.134	0.118	0.057	0.072	0.055	0.091	0.124	0.113
Brass Hill (BHill)[1]	0.094	0.109	0.083	0.096	0.107	0.072	0.094	0.110	0.085	0.096	0.104	0.069	0.030	0.033	0.023	0.082	0.085	0.049
Preston Bennet (PB)	0.724	0.701	0.370	0.765	0.739	0.371	0.710	0.691	0.369	0.532	0.564	0.409	0.103	0.189	0.235	0.761	0.785	0.267
Preston Hill I (PH_1)	0.414	0.521	0.490	0.380	0.528	0.496	0.418	0.520	0.489	0.426	0.512	0.473	0.139	0.180	0.153	0.343	0.383	0.201
Preston-Hill 2 (PH_2)	0.412	0.508	0.456	0.401	0.513	0.463	0.439	0.507	0.454	0.424	0.500	0.439	0.179	0.208	0.158	0.372	0.374	0.202
Preston Lahiri No 1 (PL_1)	0.486	35.550	4655.569	0.492	7.233	101.118	0.483	42.630	5204.835	0.487	7.201	69.995	0.148	0.155	0.066	0.478	4.790	78.060
Preston-Lahiri No 2(PL_2)	0.450	13.137	640.044	0.448	19.491	974.494	0.431	11.548	524.100	0.449	3.116	14.227	0.207	0.221	0.116	0.413	12.901	323.772
N	31,500			6,300			25,200			700			4,320			10,368		

SD, standard deviation; Med, median.

* $\theta_1 = \theta_3 = 0$

• $C_1 = C_2$ and $C_3 < 1$

‡ $C_1 \neq C_2$ and $C_3 < 1$ and maxabs$(C_1 - C_2) < .10$

[1]Values of errors in the Brass-Hill shown in the first row correspond to errors associated with the ratio C_1/C_2. While values of Brass-Hill in the seventh row correspond to erros associated with relative completeness of death registration.

[2]BMartin is a variant of Brass classic method that relaxes the assumption of stability and assumes instead past mortality decline.

Scenario 2: $C_1 = 0.85$ if age [15-35], $C_1 = 0.75$ elsewhere; $C_2 = 0.95$ if age [15-35], $C_1 = 0.85$ elsewhere; $C_3 = 0.85$ if age [15-35], $C_1 = 0.80$ elsewhere.

Search for an optimal estimate is carried out considering all prior information available and the following are general rules:

i. In the absence of any knowledge whatsoever about errors or deviations from stability, the search for best method should be concentrated on the pooled sample subset in Tables 9–11, panel A.

ii. When exogenous information suggests stability and not much else, the search should focus on the subset of stable populations in Tables 9–11, panel B. Instead, when there is prior empirical data confirming violation of stability, for example past shifts in fertility regime, but one can be agnostic about completeness and age misreporting, the search of optimal method should concentrate on the population subset in Tables 9–11, panel C.

iii. When in addition to lack of stability there is evidence of defective coverage of population and death counts but no suggestion of significant net age overstatement at adult ages, the search should shift to the subset in Tables 9–11, panel D.

iv. When the researcher suspect a scenario like in (iii) above but, in addition, there is evidence of age misreporting, identification of optimal method should be done using Tables 9–11, panel E.

v. Finally, in cases scenario (iv) is most reasonable and one can establish that completeness of two censuses is (possibly) defective but equal in both censuses, identification of the optimal choice must be done with Tables 9–11, panel F.

The results displayed in Tables 9–11, panels A through F contain a number of salient characteristics. First, as already suggested in the work by Hill and colleagues, Brass's methods to estimate relative completeness of the two censuses is uniformly good, regardless of population subset. Second, with the exception of Brass methods, the magnitude of errors are larger when census coverage is defective as long as completeness is NOT the same in both censuses. This is because all methods except Brass's rely on direct computations of age specific growth rates from the observed data, a quantity that will be in error when there is different coverage errors in two successive censuses. Indeed, the performance of these methods improves substantially when there is accurate census coverage or, equivalently, *when coverage is the same in both censuses* (Table 9, panel D). Fourth, age misreporting affects the accuracy of all estimates but substantially more so in some cases (Brass's methods and the second variant of Preston-Hill) than in others (Bennett-Horiuchi all variants). Fifth, the magnitude of errors obtain when relative completeness is age dependent (last two columns of panels A-F in Tables 9–10) varies sharply by technique but, in general, are lowest in the method by Bennett-Horiuchi.

The most important inference from this evaluation exercise is as follows: if one excludes population subsets with defective census completeness, the optimal choice is always one of the variants of Bennett-Horiuchi method followed by the two methods proposed by Brass, irrespective of violations of stability assumptions or age misreporting. This suggests the following strategies:

i. In the absence of exogenous information about the difference in completeness between the two census and if the assumption of age invariant completeness holds, use Brass method;

ii. In the absence of exogenous information, whether or not age dependence of relative completeness is suspected, use a two stage procedure: first estimate relative completeness of census enumeration using Brass' method, adjust intercensal rates of growth and then apply Bennett-Horiuchi method.

We use both strategies in LAC and when the difference between estimates was less than 0.05 we compute the average of Brass and Bennett Horiuchi estimates. When their difference exceeded 0.05 we chose the estimate from strategy (ii)[20].

6.2 Defective age reporting

Do the procedures to identify and adjust for age misreporting produce robust estimates of the true population parameters? To answer this question we select the subset of simulated populations with age misreporting and defective completeness, adjusted for completeness following strategy (ii) above, we identify the existence of age misreporting, and then correct for it using techniques (ii) in section 4.2. Tables 5 through 7 display the main results. First, Table 5 contains parameters associated with expression (4.5) and reveals that the fit is almost perfect and that the estimated constant is unit, as it should be. Table 6 shows that when the procedure is reversed and we regress cmR_x on the vectors $\alpha_{1x=45,100}$ and $\alpha_{2x=45,100}$ the errors of estimates are trifle. This suggests that if an observed population belongs to the space of simulated populations, we can retrieve estimates of the magnitude of age net over-reporting that are highly accurate by simply using the estimated relation between the observed cmR_x and estimates $\alpha_{1x=45,100}$ and $\alpha_{2x=45,100}$ from the simulated populations.

7 Discussion: the issue of uncertainty

By an large the methods to adjust mortality statistics reviewed here perform satisfactorily provided the key assumptions on which they rest are concordant with the empirical conditions that produce the data. This is most unlikely to be the case always or even frequently for one single assumption and much less for combinations of assumptions. The conventional strategy has invariably been to scrutinize alternative estimates and then settle for one based on explicit or, more frequently, implicit reasoning and judgments about concordance of assumptions and observables. We believe we can improve upon this practice.[21]

The evaluation study generates a superpopulation of errors associated with the application of each technique under conditions that violate to different degrees one or several of the cardinal assumptions on which they rely. It follows that for each technique we can define precisely the magnitude of error—however measured— associated with conditions that depart from the combination of assumptions in *ex ante* known ways. In our simulation the base universe of populations was generated by combining different demographic parameters (levels and patterns of fertility and mortality) thus producing multiple instances where one could alter conditions imparting changes that violate assumptions(lack of stability, adult migration, variable completeness, age misreporting that departs from assumed patterns etc.). As a consequence, we have all the information needed to define the frequency distribution of errors associated with one technique under one set of simulated conditions. And, in particular, one can define the probability that a singular technique will

[20]It is important to note that when relative completeness is age dependent, Bennett-Horiuchi is *mean optimal*, in the sense that the weighted average of relative death completeness of observed data will be best estimated by Bennett-Horiuchi methods. It does not mean that, once applied, the adjusted mortality rates (and derived function of the life table) will also be best estimates. None of the methods we include in our evaluation can escape from the assumption of constant relative completeness and, therefore, we can only aspire to find a mean optimal candidate.

[21]An application of the ideas described here is in Palloni and Beltrán-Sánchez (2016).

produce an error less than, say 5, percent under a given set of well defined (simulated) conditions that possibly depart from assumptions.

Now, assume that in any population we know the probability that the historical conditions that produced the data match one of the multiple sets of simulated conditions.[22] It would then be possible to compute the unconditional probability that, in that particular population, a given technique will produce an error of less than 5 percent. If one repeats this for all candidate techniques than can be deployed, we will have alternative values of the true parameters and known magnitude of uncertainty associated with each of them. This is sufficient knowledge to analyze the data incorporating uncertainty-rather than ignoring it by a sleigh of hand choosing the value discerned, however convincingly, to be the true parameter. The outcome of this is that target parameters such as the rate of decline of mortality rates for a given age group, the effects of income on mortality changes, the fraction of life expectancy improvements associated with income changes, will be associated with bounds of uncertainty and the standard errors of these estimates will fall within a range rather than being point estimates. This may be less pleasing than providing a single value (with associated standard errors) but it is also a strategy that fully admits levels of ignorance. Since some of the estimates could be used for projections and forecasts, it stands to reason that the above procedure will lead to probabilistic forecasts by virtue of uncertainty of estimates not just because of uncertainty about future trends.

[22]These probabilities can be constructed from expert judgments.

References

Bhat, M. P. (1987), "Mortality in India: Levels, Trends and Patterns," Ph.d., University of Pennsylvania.

Bhat, P. N. M. (1990), "Estimating Transition Probabilities of Age Misstatement," *Demography*, 27, 149–163.

Brass, W. (1975), *Methods for Estimating Fertility and Mortality from Limited and Defective Data: based on seminars held 16-24 September 1971 at the Centro Latinoamericano de Demografía (CELADE) San José, Costa Rica*, Chapel Hill, NC: International Program of Laboratories for Population Statistics, Dept. of Biostatistics, School of Public Health, Carolina Population Center, Univ. of North Carolina at Chapel Hill.

Chidambaram, V. and Sathar, Z. (1984), "Age and date reporting," in *Comparative Studies: Cross-National Summaries*, Voorburg, Netherlands: International Statistical Institute, vol. 5.

Coale, A. J., Demeny, P., and Vaughan, B. (1983), "West? Model Life Tables and Stable Populations," in *Regional Model Life Tables and Stable Populations (Second Edition)*, ed. Vaughan, A. J. C. D., Academic Press, pp. 41–154.

Coale, A. J. and Kisker, E. E. (1986), "Mortality Crossovers: Reality or Bad Data?" *Population Studies*, 40, 389–401.

Condran, G. A., Himes, C., 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.

Dechter, A. R. and Preston, S. H. (1991), "Age misreporting and its effects on adult mortality estimates in Latin America," *Popul Bull UN*, 31-32, 1–16.

Del Popolo, F. (2000), "Los Problemas en la Declaración de la Edad de la Población Adulta Mayor en los Censos," Report, CELADE, CEPAL, ECLAC.

Elo, I. T. and Preston, S. H. (1994), "Estimating African-American Mortality from Inaccurate Data," *Demography*, 31, 427–458.

Ewbank, D. C. (1981), *Age Misreporting and age-selection and Remuneration: Sources, Patterns, and Consequences for Demographics Analysis*, Committee on Population and Demography, Washington DC: National Reserach Coucil.

Grushka, C. O. (1996), "Adult and old age mortality in Latin America: Evaluation, adjustments and a debate over a distinct pattern," Thesis, University of Pennsylvania.

Hill, K. (1987), "Estimating Census and Death Registration Completeness," *Asian and Pacific Population Forum*, 1, 8,23–13,24.

— (2003), "Adult mortality in the developing world: what we know and how we know it," .

Hill, K. and Choi, Y. (2004), "Performance of GGB and SEG given various simulated data errors," .

Hill, K., Choi, Y., and Timaeus, I. (2005), "Unconventional approaches to mortality estimation," *Demographic Research*, S4, 281–300.

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.

Horiuchi, S. and Coale, A. J. (1985), "Age Patterns of Mortality for Older Women," .

Kamps E., J. (1976), *La declaración de la edad en los Censos de Población de la América Latina; exactitud y preferencia de dígitos en los Censos de 1950, 1960 y 1970*, vol. Serie C of *Series Históricas*, San José, Costa Rica: ONU, CEPAL, CELADE.

Martin, L. (1980), "A Modificationfor Use in Destabilized Populations of Brass's Technqiue for Estimating Completeness of Death Registration," *Population Studies*, 34, 381–95.

Mazess, R. B. and Forman, S. H. (1979), "Longevity and age exaggeration in Vilcabamba, Ecuador," *J Gerontol*, 34, 94–8.

Nuñez, L. (1984), "Una Aproximacion al Efecto de la Mala Declaracion de la Edad en la Informacion Demografica Recabada en Mexico," Report, Direccion General del Registro Nacional de Poblacion e Identificacion Personal.

Ortega, A. and Garcia, V. (1985), "Estudio Sobre la Mortalidad y Algunas Caracteristicas Socioeconomicas de las Personas de la Tercera Edad: Informe de la Investigacion Efectuada en los Cantones de Puriscal y Coronado del 3 al 20 de Junio de 1985," Report, CELADE.

Palloni, A. (1990), "Assessing the Levels and Impact of Mortality in Crisis Situations," in *Measurement and Analysis of Mortality: New Approaches*, eds. Vallin, J., D'Souza, S., and Palloni, A., New York: Oxford University Press, pp. 215–49.

Palloni, A. and Beltrán-Sánchez, H. (2016), "Uncertainty and Mortality Estimates in the Latin American and Caribbean (LAC) Region: 1850–2010," *working paper*.

Palloni, A. and Pinto, G. (2000), "Fertility patterns and change in Latin America," .

— (2004), "One hundred years of mortality in Latin America and the Caribbean: the fragile path from hunger to longevity," Presented at the annual meeting of the Population Association of America, Boston, Massachusetts, April 1-3, 2004.

Preston, S. and Hill, K. (1980), "Estimating the Completeness of Death Registration," *Population Studies*, 34, 349–366.

Preston, S. H. and Bennett, N. G. (1983), "A Census-based Method for Estimating Adult Mortality," *Population Studies*, 37, 91–104.

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, 193–209.

Preston, S. H., Elo, T. I., Hill, M. E., and Rosenwaike, I. (2003), *The Demography of African Americans, 1930-1990*, Boston: Kluwer Academic Publishers.

Preston, S. H. and Lahiri, S. (1991), “A short-cut method for estimating death registration completeness in destabilized populations,” *Math Popul Stud*, 3, 39–51.

Rosenwaike, I. (1987), “Mortality Differentials among Persons Born in Cuba, Mexico and Puerto Rico Residing in the United States, 1979-81,” *American Journal of Public Health*, 77, 603–606.

Rosenwaike, I. and Preston, S. H. (1984), “Age Overstatement and Puerto Rican Longevity,” *Human Biology*, 56, 503–525.

Spencer, G. (1984), “Mortality among the Elderly Spanish Surnamed Population in the Medicare Files: 1968 to 1979,” .

A Appendix. Definition of demographic profiles for the simulation

Five different master populations were created, one stable and four nonstable populations. In each case we start with a stable population in 1900 and we compute yearly populations until the year 2000. The age distribution is in single years of age but for totals (not by gender).

The four non-stable populations were generated following approximately the mortality and fertility schedules for Costa Rica, Mexico, Guatemala and Argentina, Uruguay for the period 1900-2000.

A.1 Stable population

The stable population is generated using constant values for $GRR = 3.03$ and $E(0) = 45$ for the period 1900 and 2000 with a natural rate of increase $r = 0.025$.

A.2 Non-stable populations (a)(b)(c)

	I			II			III			IV		
Year	$E(0)$	GRR	r	$E(0)$	GRR	r	$E(0)$	GRR	r	$E(0)$	GRR	r
1900	34.70	3.60	0.05	26.30	6.20	0.04	22.10	5.80	0.03	45.40	1.80	0.02
1910	35.10	3.40	0.05	29.60	5.70	0.04	25.40	5.70	0.03	48.90	1.70	0.02
1920	35.10	3.20	0.05	32.90	5.20	0.04	28.70	5.20	0.03	51.30	1.60	0.02
1930	42.20	2.60	0.05	36.20	4.70	0.04	32.00	4.70	0.03	54.40	1.50	0.02
1940	46.90	2.50	0.05	41.80	4.20	0.04	37.40	3.80	0.03	59.60	1.40	0.02
1950	55.60	2.40	0.05	50.70	3.40	0.04	40.20	3.50	0.03	66.30	1.30	0.02
1960	62.60	2.30	0.05	58.50	3.30	0.04	47.00	3.30	0.03	68.40	1.40	0.02
1970	65.40	2.10	0.05	62.60	3.20	0.04	53.90	3.10	0.03	68.80	1.50	0.02
1980	72.60	1.70	0.05	67.70	2.10	0.04	58.20	3.00	0.03	71.00	1.30	0.02
1990	75.70	1.50	0.05	71.50	1.50	0.04	62.60	2.60	0.03	72.80	1.20	0.02
2000	77.30	1.30	0.05	73.40	1.20	0.04	65.90	2.20	0.03	75.20	1.10	0.02

(a) Non Stable population I, II, III and IV follow the patterns of mortality and fertility between 1900 and 2000 assessed with current (Adjusted data) for Costa Rica, Mexico, Guatemala and Argentina/Uruguay respectively.

(b) Population parameters were directly estimated for each decade and then interpolated linearly within each decade to obtain yearly values.

(c) The initial population age distribution for I, II and III correspond to the stable population associated with parameter values in 1900. In case IV the initial population corresponded to the average of census populations closest to 1900.

B Appendix. Proof of lack of identification of parameters of net age overstatement

Using the same notation as in the text we have

$$\Pi^T = (1/\phi^{no})[\hat{\Theta}^S]^{-1}\Pi^O$$

and

$$\Delta^T = (1/\lambda^{no})[\hat{\Theta}^S]^{-1}\Delta^O.$$

In a closed population the relation between the vectors for populations in two successive censuses and the vector of intercensal deaths is:

$$\Pi^T_{t+k} = \Pi^T_t + \Delta^T_{[t,t+k]} \,. \tag{B.1}$$

Using the first two expressions in (B.1) yields:

$$(1/\phi^{no})[\hat{\Theta}^S]^{-1}\Pi^O_{t+k} = (1/\phi^{no})[\hat{\Theta}^S]^{-1}\Pi^O_t - (1/\lambda^{no})[\hat{\Theta}^S]^{-1}\Delta^O_{[t,t+k]} \,. \tag{B.2}$$

From (B.2) we see that only (ϕ^{no}/λ^{no}) is identifiable with the available information.

C Appendix. Behavior of the age misreporting index $cmR^o_{x,[t_1,t_2]}$

The expression of the age misreporting index is

$$cmR^o_{x,[t_1,t_2]} = \frac{cmP^o_{x+k,t_2}/cmP^o_{x,t_1}}{1-(cmD^o_{x,[t_1,t_2]}/cmP^o_{x,t_1})}$$

a ratio of two different estimators of the same quantity, namely the cumulative probability of survival of the population aged x and over at time t_1 to age $(x+k)$ and over at time t_2. Use of cumulative quantities in the index is an important prerequisite since it minimizes the impact of age misreporting within the bounds of the cumulative quantities. Thus, erroneous transfers over age x do not affect population counts at ages x and over. These quantities are influenced only by transfers from ages younger than x into ages x and above or by transfers from ages x and above to ages younger than x. Admittedly, however, use of cumulative quantities complicates the algebra and muddles interpretation. To circumvent this difficulty and preserving the same set up and assumptions defined in the text, we redefine the expression for single years of age to obtain:

$$R^o_{x,[t_1,t_2]} = \frac{P^o_{x+k,t_2}/P^o_{x,t_1}}{1-(D_{x,[t_1,t_2]}/P^o_{x,t_1})}$$

or the ratio of a conventional survival ratio computed from two successive population counts to the survival ratio computed from the complement of a measure of the conditional probability of dying between the two censuses. If the population is stationary, the numerator is simply the ratio L_{x+k}/L_x in a life table and the denominator is the complement of the probability of dying in the intercensal period, namely, $1-(1-L_{x+k}/L_x)$. From this it follows that,

$$\ln\left(R^o_{x,[t_1,t_2]}\right) \sim -I^N_{x,x+k} - \ln\left(1-\left[1-exp\left(-I^D_{x,x+k}\right)\right]\right) \tag{C.1}$$

where $I^D_{x,x+k}$ and $I^N_{x,x+k}$ are estimators of the integrated hazards between x and $x+k$ consistent with the survival ratios in the denominator and numerator respectively. When the population is closed to migration, there is perfect coverage and no net age overstatement, expression (C.1) equals 0 as both estimators of the integrated hazards are identical. When there is age overstatement expression (C.1) becomes

$$\ln\left(R^o_{x,[t_1,t_2]}\right) \sim \ln\left(\frac{h(x+k)}{h(x)}\right) - I^N_{x,x+k} - \ln\left(1-\frac{g(x)}{h(x)}\left[1-exp(-I^D_{x,x+k})\right]\right) \tag{C.2}$$

where $h(.)$ and $g(.)$ are defined in the text and refer to increasing functions of age that reflect age overstatement of ages of population and deaths respectively. When these functions are equal to

1, there is neither population nor death age overstatement or, if there is, their effects cancel each other out. Expression (C.2) can be simplified if we expand the inner log expression in a Taylor series around a value of $f(x) = g(x)/h(x) = 1$:

$$\ln\left(R^o_{x,[t_1,t_2]}\right) \sim \ln\left(\frac{h(x+k)}{h(x)}\right) - I^N_{x,x+k} + \left(\frac{g(x)}{h(x)} - 1\right)(1 + I^D_{x,x+k}) + I^D_{x,x+k} \tag{C.3}$$

an expression that reduces to 0 when $h(x+k)/h(x)$ =1 and $f(x) = 1$.

Expression (C.3) is the analytic support for inferences regarding the effects of age overstatement on the index of age misstatement $cmR_{x,[t_1,t_2]}$ (see text). Deviations from the assumption of population stationarity introduce only minor changes in the algebra but leave the implications of expression (C.3) intact. However,when, as required by the original index, we restore the cumulative functions, the algebra becomes intractable even in the case of a stationary population. The way out of this conundrum is to think of the cumulative ratios as functions not of the exact integrated hazards, as in expressions (C.1)-(C.3) but rather as expressions of mean values of corresponding integrated hazards. Thus, in a stationary population, the survival ratio of the cumulative populations at ages x and $x+k$ is the ratio $T(x+k)/T(x)$ which can be written as $\int_{x+k}^{\infty}[exp(-\int_0^y \mu(s)ds)]dx / \int_x^{\infty}[exp(-\int_0^y \mu(s)ds)]dx$. Using the mean value theorem in numerator and denominator leads to the approximation $exp(-\int_{x+i}^{x+k+i'} \mu(s)ds)$ or, more generally, $exp(-\int_{x^*}^{x^{**}} \mu(s)ds)$ where $x^* > x$ and $x^{**} > x+k$. Upon taking logs in this expression we retrieve an integrated hazard that expresses integration of the force of mortality over two ages that are not fixed ex ante (such as x and $x+k$) but, rather, between limits (ages) that are a function of the underlying force of mortality. For this reason, in the text, we use the symbols $I^N_{x,x+k}$ and $I^D_{x,x+k}$ associated with cumulative quantities as "integrated hazard analogues".