Appendix. Data Construction and Sources

All estimates of life expectancy and its dispersion in this paper come from two kinds of related data: life tables or mortality rates. A life table lists values of the survival function

$$S(T) = 100{,}000 \prod_{t=0}^{T} s(t) \qquad \text{(A1)}$$

where,

S (T) = number of survivors to the beginning of year T out of a cohort of 100,000 births at the start of t=0

s (t) = probability that an individual will survive from the beginning to the end of year t

That is, the life table shows how many are expected to survive to each age out of hypothetical birth cohort of 100,000. The function is assumed to go to zero at an upper limit which is usually around 110 years. Mortality rates give the probability of dying in the next year as a function of age, and this is related to life table data because

$$s(t) = 1 - m(t) \qquad \text{(A2)}$$

where,

m (t) = probability of dying over the next year for an individual t years old

So, the life table can be thought of as giving a stock of survivors at each age that depreciates at the mortality rate for that age. There are some conceptual and practical differences between life table and mortality measures, but for present purposes I use the two interchangeably.[1]

The most commonly available life table is the period life table, which is based on contemporaneous mortality experience. I use period life tables exclusively here. The less common cohort life table follows a birth cohort through its subsequent history. The dominant reason for using period life tables is practical: a reasonably complete cohort life table of today would be tracking a cohort born in the early 20th century, so it would discard much of the available mortality information for subsequent birth cohorts.

While period life tables are often described in forward-looking terminology, they are really a snapshot of current mortality experience. Thus, the oft-cited life expectancy at birth statistic is not an estimate of how long infants born today can reasonably expect to live. It is just a summary statistic of current mortality experience. Specifically, expected life is the mortality weighted average age of death for the current population. This is an expected value for a counterfactual in which infants born today will experience the same mortality risks over their lifetimes as do the various age groups in the current population. Thus there is no allowance for medical progress, which is a historically important shortcoming of life expectancy measures if they are treated as predictions.

[1] For example, mortality rates are measured as deaths of individuals of age X in calendar year Y divided by estimated population of age X on June 30. Life tables are supposed to give the probability of dying by December 31 of Y for those alive on January 1. The simple relation in (A2) may not hold if deaths are bunched within the year. This is the case with infant mortality, where most deaths occur within a few days of birth. For this reason, mortality data are often adjusted for purposes of constructing life tables, and there may be a discrepancy between published mortality rates and the rates implied by published life tables.

A. International Comparisons

The data for the cross-country samples come from online databases that cover many countries and from print and electronic sources specific to single countries. These sources are listed in Table A1. I took available data for the country/years in my sample at face value without any attempt at 'quality control.' Accordingly, the data I used span a variety of methods, from estimates by demographers based on a few data points to summaries of all recorded deaths. My sample does not include every country I found in the literature. A necessary condition for including a country was that I could construct a reasonably complete time series for the 20th century. A sufficient condition was that I could start such a series no later than the mid 19th century forward. The ten countries which met the sufficient condition are sometimes referred to as the 'long' sample here and in the main text. (Half of the long sample is from Scandinavia, where 18th and 19th century church records were unusually detailed.) I included the other countries (the 'short' sample) based mainly on their size and how far back before 1900 I could begin their time series.

Table A2 lists the countries in each sample, the start date of their time series, data sources and details on the handling of the missing value problems described below.

Each time series used in the paper – such as the mean and dispersion of lifetimes in a specific country- is derived from a corresponding historical inventory of life tables. For purposes of comparison and aggregation, these inventories were sorted into 5 year groups beginning 1740-44 and ending 2000-04. Each group of years is identified in the graphs and other analyses by its mid-point 1742, 1747… 2002. Complete life tables use one-year age intervals. That is, in terms of equation (A1), they give S (0), S (1)… S (110). These are denoted in the

literature as 1 x K life tables, where K is the number of years of mortality experience used to estimate each of the points on S (t). K is often set greater than 1 to reduce the measurement error inherent in using a single year's mortality within a one-year age interval. Wherever feasible I used or estimated a 1 x 5 life table and assigned it to the year group closest to the mid-point of the life table.[2]

However, it was not always feasible to use a 1 x 5 life table. And there were gaps in the historical record. Accordingly there are two kinds of missing data problems, which I treated as follows:

1. Incomplete or "abbreviated" life tables, or equivalently and more commonly, mortality rates over an age interval. Most available mortality rate data cover 5 or 10 year age intervals. So they yield only a few points on S (t). A typical example might give S (1), S (5), S (10), S (20)…S (80). When I encountered such data I converted them to complete life tables. To do this I assumed constant mortality rates within any age interval from 5 to 65. This assumption imparts some bias,[3] but it is inconsequential empirically. The consequential problems arise at very young and very old ages. Historically the vast majority of early childhood mortality occurs in the first year. So there can be considerable distortion, especially of dispersion measures, from assuming constant mortality over the early childhood years. At the other end, there is a distinct acceleration of mortality risk beyond age 70, so values of S (70+) are not well approximated by a constant mortality rate in old age.

[2] The common alternative to a 1 x 5 life table is a 1 x 1. I assigned those to the closest matching 5 year age group. Occasionally I encountered 1 x 10 life tables. Here I assumed the table is valid for each 5 year sub-period and proceeded as described.

[3] For example, mortality beyond age 50 or so rises palpably with age. So assuming constant mortality from, say, 50 to 60 will underestimate points like S (55), or overestimate mortality rates at ages below 55.

In cases where I had to estimate the values of S (t) for early childhood or old age I used a contemporary 'reference country' which had a complete life table available and an otherwise similar mortality profile. I filled out the subject country's life table by assuming that the share of deaths at each specific age in an interval was the same as the reference country's share.[4]

2. Gaps in time series arose when a country is missing life table or mortality data that fit within one or more of the 5 year periods between the beginning and end of the series. I filled gaps of 15 years or less by linear interpolation of the terminal S () values. For longer gaps, I used a ratio-to-trend method. This entailed selecting a reference country with data available within the gap. Then I interpolated the subject and reference country's values within the gap. The estimated value for the subject country is then set to the subject county's interpolated value times the ratio of the reference country's actual value to its interpolated value.

Column (3) of Table A2 indicates which countries required interpolation of either life table or time series data and lists any reference countries. In general, the less developed countries in the short sample have the less complete data, and the consequently heavier use of interpolation needs to be kept in mind when evaluating their data.

Wars often create considerable discontinuities in a country's time series. However, I left these discontinuities in much of the data (e.g., figures 2 through 4), because their relatively small size lends some emphasis to the importance of longer run trends.[5] However, I did remove war related effects from the analysis of gender inequality, because they often obscure everything else.

[4] I found that the choice of reference country does not much affect the summary measures that I ultimately used. Most any contemporary life table will have a similar age distribution of mortality in the tails of the age distribution.
[5] However, the figures obscure some of the effects of war due to interpolation around gaps in the data that occur in war time periods.

I did this by interpolating all the gender specific measures around war-affected periods. Column (4) of Table A2 gives the mid-point of any war-affected periods for each country.[6]

B. US States and Counties

a. States, 1900-2000

Data for US states come from life tables constructed from mortality rates. These are available each ten years for a growing subset of US states from 1900 to 1930, and then for all states thereafter.[7] The mortality rates up to 1960 come from:

Linder, Forrest E. and Robert D. Grove (1943).*Vital Statistics Rates in the United States, 1900-1940.* Washington: Government Printing Office

and

Grove, Robert D. and Alice Hetzel (1968). *Vital Statistics Rates in the United States, 1940-1960.* Washington: Government Printing Office

These can also be found at

http://www.cdc.gov/nchs/datawh/statab/unpubd/mortabs/hist290.htm

These were converted to 1 x 1 life tables by filling in specific ages from the contemporary US life table (see previous section). I assumed that the share of deaths at a specific age within an age interval was the same for each state as the corresponding share from the US life table.

[6] The absence of a listing in this column need not imply that the country was unaffected by war, because data for that country may be missing (and therefore already estimated by interpolation). For example, German data for 1917 is missing and therefore estimated by interpolation between 1912 and 1922. Accordingly no further adjustment for the effects of World War I is necessary for that country.

[7] There are some state-specific life tables for the early period, but they cover fewer states and sometimes only the white population. I opted for consistency and breadth here and accordingly constructed all the state life tables over the 20th century from mortality data.

The incomplete coverage arises because the US did not adopt the internationally uniform death registration system until 1932. Prior to that, some states implemented registration systems at different times. Thus in 1900 there were 9 registration states and by 1920 there were 35. The states adopting earliest tended to be the larger and more industrialized states.

For post-1960 data we have state specific 1 x 3 life tables around each census year in

US Center for Disease Control. National Center for Health Statistics. *US Decennial Life Tables. State Life Tables* (various issues).

These can also be found at

http://www.cdc.gov/nchs/products/pubs/pubd/lftbls/lftbls.htm

b. Counties, 1970-2000

I use county mortality data for the early 1970s, 1980s and 2000s. (I leave out the early 1990s to avoid distortion from the AIDS epidemic that begins in the early part of the 1980s and crests in the mid 1990s.) I then converted these data to 1 x 5 life tables for each county/period. For each of the three decades I collected five years of mortality rates for each US county centered on 1972, 1982 and 2002 respectively. The data for the 1980s and 2000 county mortality rates are available from

http://wonder.cdc.gov/

For 1970 I combined mortality counts by county with census county population estimates to produce the mortality rates for that year. The number of deaths is available from the compressed mortality file for the relevant years as described at

http://www.cdc.gov/nchs/products/elec_prods/subject/mcompres.htm

This is a summary of individual death certificates which includes an indicator of the decedent's county of residence and age. County population estimates by five year age group were taken from the US census website at

http://www.census.gov/popest/archives/pre-1980/co-asr-7079.html

The individual death certificates for the 1970s were sorted into age group/county bins, counted and converted to rates by dividing by the population estimates.[8]

All of the county mortality rates were converted to 1 x 5 life tables by using the contemporary US life table to fill in age intervals and allocate ages beyond 85.[9] The mortality variables are then constructed from these life tables in the same way as described previously.

The socio-economic variables used in the analysis of the county data come from 1970-2000 census data accessed through the National Historical Geographic Information System (NHGIS) at

http://nhgis.org/

The available income and education distributions were treated as follows:

Education. The Census distributions of completed education by the population over 25 use varying definitions of the relevant intervals. For example, after 1980 the distribution of college

[8] We do not have county population detail for ages less than 5. To estimate the infant mortality rate (and the 1 to 4 rate) I allocated each county's under-5 population according to state shares by age and then used these estimates as the denominator of the relevant mortality rate.

[9] Because of coding problems, Alaska was treated as one county and some areas in Virginia were dropped. For some small counties data are occasionally missing for specific years or for specific age groups. In these cases I filled gaps from the relevant average of small counties in the same state

attendees is given by degree attained rather than by years completed. I converted each distribution to a 0 to 21 year scale in a two step procedure. First I constructed a state distribution of completed years by using available information from adjacent years to allocate specific years. This is straightforward for years up through 12. For example, the 1970 distribution separates 5 to 6 years from 7, while the 1980 distribution does not. Here I allocate the 1980 5 to 7 distribution into 5 to 6 and 7 years according to the available 1970 shares. When separate years could not be estimated (as is the case here with 5 to 6 years), I allocate the shares equally to each year.

For years beyond high school, where we sometimes have years and sometimes degrees, I match degrees and years beyond 12 as follows: Associates degrees 2 years, Bachelors 4 years, Masters split evenly between 5 and 6 years, Professional 7 years and Doctorate 9 years. Then I use the 1990 shares to allocate the more aggregated 1980 and 1970 data for college attendees. These adjustments are somewhat arbitrary, but the lumpy character of education distributions renders their effects inconsequential. Typically there is a large mass at 12 years, another clump at 16, a sizeable group between the two and few in the long tail beyond 16 years.[10]

The state distributions are then used to allocate each county's distribution to specific years, and I extract the county mean and standard deviation of the resulting distribution.

Income. I use family income distributions, which have open-ended upper tails. To close the upper tail I fit a Pareto distribution to each county's income distribution using the upper two income classes to estimate the parameter of this distribution and then the implied mean income

[10] For example in 2000, 34 per cent of the population has exactly 12 years, another 25 per cent have less than 4 years of college and 13 per cent have a 4 year degree. Of those 7 per cent who get beyond 4 years of college only 1 in 10 have a doctorate.

in the upper tail. I truncate the parameter at the 99th percentile of the sample distribution to avoid outliers, and I use the sample mean where the parameter cannot be estimated.[11] I then assume that all families earn either the mid-point of their distribution or the mean of the upper tail. The mean, mean of logs and its standard deviation are then extracted from this distribution.

[11] As when there is mass in the upper tail but none in the penultimate class.

Table A1. Sources of Data

[Brief reference used in Table A2 in brackets]

A. Online Databases Covering Multiple Countries.

[HMD] Human Mortality Database. (http://www.mortality.org/, http://www.lifetable.de/) This is a joint project of the Department of Demography, University of California, Berkeley and the Max Planck Institute for Demographic Research, Rostock, Germany. Each site has downloadable life tables for a number of countries and years. Wherever available I use the 1 x 5 (1 year age interval, 5 year average data) period life table.

[WHO] World Health Organization. WHO Statistical Information System (WHOSIS) (http://www.who.int/whosis/en/) Mortality rates for member countries for various years from 1979 to date and abbreviated annual life tables for member countries from 1999 to date.

B. Specific Countries

[AR1] Somoza, Jorge (1971). La Mortalidad en la Argentina Entre 1869 y 1960. Centro de Investigaciones Sociales. Instituto Tocuarto Di Tella. Centro Latinoamericano de Demografia. Editorial del Instituto.

[AR2] Centro Latinoamericano y Caribeño de Demografía- División de Población (CEPAL/CELADE), Boletín Demográfico (No.67 de enero del 2001). Web site: http://www.eclac.cl/publicaciones/Poblacion/9/LCG2119P/BD67.html

[AU1] Australian Bureau of Statistics. Cat. no. 3105.0.65.001 Australian Historical Population Statistics. Table 50. Number of Persons at Exact Age x, Australia 1881 onwards. Available at: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3105.0.65.0012006?OpenDocument

[BR1] Arriaga Eduardo. (1968). New life tables for Latin American Populations in the Nineteenth and Twentieth Centuries. Population Monograph series No. 3. University of California Berkeley. Department of Demography.

[BR2] Instituto Brasileiro de Geografia e Estatistica (IBGE), Diretoria de Pesquisas, Departamento de População e Indicadores Sociais. Indicadores Demograficos (various issues).

[CL1] Preston, Samuel; Nathan Keyfitz and Schoen Robert. (1972). Causes of Death. Life tables for National Populations. Seminar Press. New York and London.

[CL2] United Nations. Statistical Office. Demographic Yearbook (various issues). New York

[FI1] Turpeinen, Oiva. Fertility and Mortality in Finland since 1750. *Population Studies*. Vol 33 N 1. March 1979. 101-114

[GE1] Imhof, Arthur. (1990). Lebenserwartungen in Deutschland vom 17. bis 19. Jahrhundert. (Life expectancies in Germany from the 17th to the 19th Century). Acta Humaniora. Weinheim.

[IN1] S. P Jain. (1981) Actuarial Report and Life Tables 1951-1961. Office of the Registrar General, Ministry of Home Affairs, New Delhi

[IN2] Compendium of India's Fertility and Mortality Indicators (1971-1997) Based on the Sample Registration System. Registrar General. Ministry of Home Affairs. New Delhi. 1999

[IN3] Sample Registration System, 1970-1975. Vital Statistics Division, Office of the Registrar General, Ministry of Home Affairs, New Delhi, 1983

[IN4] Sample Registration System, 1979-1980. Vital Statistics Division, Office of the Registrar General, Ministry of Home Affairs, New Delhi, 1984

[IN5] SRS Abridged Life Tables 1981-1985. Vital Statistics Division, Office of the Registrar General, Ministry of Home Affairs, New Delhi, 1990.

[IN6] SRS abridged life tables 1986-1990. Vital Statistics Division, Office of the Registrar General, Ministry of Home Affairs, New Delhi, 1994.

[IR1] Pascua, M. “Evolution of Mortality in Europe during the Twentieth Century” in World Health Statistics Report, v. 3. World Health Organization. Geneva, 1950.

[IT1] Natale, Marcello and Amedeo Bernassola. (1973) La Mortalita per Causa Nelle Regioni Italiane. Tavole per Contemporanei 1965-66 e per Generazioni 1790-1964. Facolta di Scienze Statistiche Demografiche ed Attuarali dell' Universita de Roma. Instituto di Demografia. Roma.

[JP1] Statistics Bureau and Statistical Research and Training Institute. Ministry of Internal Affairs and Communication. Historical Statistics of Japan. Chapter 2: Population and Households. Table: 2-31- b: Complete Life Table (1891--2000). Web site: http://www.stat.go.jp/english/data/chouki/02.htm

[US1] Haines, Michael. Estimated life tables for the United States (1850-1900). NBER Series on Historical Factors in Long Run Growth. NBER Historical paper # 59. National Bureau of Economic Research, September 1994. pp 21 (US Model life tables)

Table A2. Data Sources and Methods

Country	Start Year	Years Covered	Interpolation Conventions	War Years	Sources
A. 10 Country ('long') sample	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Denmark	1837	all			HMD
England & Wales	1842	all		1912, 1917, 1942	HMD
Finland	1752	all	Sweden (LT)	1917, 1942	FI1 to 1877. Then HMD
France	1807	all		1807, 1812,1912, 1917, 1942	HMD
Iceland	1837	all			HMD
Germany	1742	1742-1852 by decade. 1872-1912,1927, 1932, 1947 on	linear, France	1942	GE1 to 1852.Then HMD
Netherlands	1852	all		1942	HMD
Norway	1847	all			HMD
Sweden	1747	all			HMD
USA	1852	1852-1892 by decade. Then 1902 on	linear		US1 to 1902.Then HMD

Table A2. Data Sources and Methods (cont.)

Country	Start Year	Years Covered	Interpolation Conventions	War Years	Sources
B. 13 country ('short') sample					
Argentina	1882	1882,1912,1917,1947 on.	Italy. Italy (LT)		AR1 to 1957. AR2 1962-97. Then WHO
Australia	1882	all		1912,1917	AU1 to 1917. Then HMD
Brazil	1872	1872,1892,1902,1922, 1942,1952,1962,1967, 1977 on.	Italy to 1902, then Spain. Linear.		BR1 to 1962. Then BR2
Chile	1907	1907,1922,1932,1942, 1952 on	Linear. Spain(LT), Portugal(LT)		CL1 to 1952. Then CL2
India	1892	1892-1907,1922,1927, 1942-1987, 1997 on	linear, India (LT)		CL2 to 1947. IN1 for 1952,1957.IN2 for 1962, 1967.IN3 for 1972. IN4 for 1977. IN5 for 1982. IN6 for1987. Then WHO
Ireland	1902	1902,1912, 1927, 1937-1952, 1957	linear, Ireland (LT)		IR1 for 1902, 1912. Then HMD
Italy	1862	all		1917,1942	IT1 to 1867, then HMD
Japan	1892	1892-1902, 1912,1922, 1927, 1937, 1947 on	linear		JP1 to 1937. Then HMD
New Zealand	1882	1882,1902,1912,1922, 1927, 1937, 1947 on	linear		CL1 to 1937. Then HMD
Portugal	1902	1902,1912,1922, 1932, 1942 on	Linear. Italy (LT), Spain (LT)		IR1 to 1947. Then HMD
Russia/Soviet Union	1897	1897,1927, 1937, 1957,1967,1977, 1982, 1987, 2002	linear, Spain		CL2 for 1897, 1927. HMD for 1937-1987. WHO for 2002. (See Note 1)
Spain	1902	all		1937,1942	HMD
Switzerland	1877	all			HMD

Table A2. Data Sources and Methods (cont.)

Start year: the mid-point of the 5 year interval with the first observation in a country's time series. (All years in the table are mid-points of 5 year intervals.)

Years Covered: years with gaps in the time series. No gaps denoted by 'all'

Interpolation Conventions: methods by which gaps in the time series and missing ages in the life tables were filled. (See appendix text for discussion). 'Linear' means that some or all gaps in the time series were filled by linear interpolation. Country name is indicated when a ratio-to-trend method is used to fill time series gaps. Country (LT) indicates which country's life tables have been used to fill missing ages in any of this country's life tables.

War Years: Years in which available data on gender differences is replaced by an interpolated value because of distortion by wars.

Sources: See Table A1 for full citations.

Notes:
1. Data are for Russian Empire (1897), Soviet Union (1927-1987), and Russia (2002). WHO data permit comparison between data for Russia and whole of former Soviet Union for 1992; they are substantially identical.