Methods protocol of the digital collection: *Causes of death — Life tables for national populations*

Version 1.0

Tim Riffe[1], Markus Göhler[1], and Adrien Remund[2,3]

[1]Max Planck Institute for Demographic Research
[2]Université de Genève
[3]Institut national d'études démographiques

May 18, 2017

1 Introduction

This data collection, originally published on paper in 1972 (Preston et al. 1972), contains mortality data from 48 countries, representing 180 different lifetables by year, sex, and population. In the original version, several data products were printed for each lifetable, including a table of death rates by twelve causes, an all-cause lifetable, a cause-decomposition of lifetable survivorship, and cause-deleted lifetable survivorship. Of these four tables, we have digitally captured, and make available only the first: all-cause and cause-specific abridged-age death rates, as well as all-cause death counts, and the mid-year population used as a denominator. This data product, which is true to the original, is available in a single long-format file, as well as in country-specific files. Furthermore, we have added value to the original data collection by graduating all rates to single ages, and extending rate schedules from age 85+ to age 100. Single-age data are also available in a single long-format file (different formats) and in country-specific files. Datasets are stored and released as both `csv` and `Rdata` files.

Data are available free of charge and without registration. This project is hosted as a satellite project of the Human Life-Table Database, and is available under the `Links` tab of its main website `http://www.lifetable.de/`. This protocol describes the original abridged and the graduation-extrapolation-modified dataset formatting. We give a synopsis of the data quality checks carried out and the graduation procedure used in order to estimate rates in single ages.

2 Format of the data

The set of digitized death rates that we now make available was originally given in wide format, with ages in rows, and death rates by cause over columns. Ages are given in standard abridged format (0,1,5,10,...,85+). Exposures are approximated using mid-year populations, and are given as integers, sometimes rounded to 1000s. All original death rates are given to five decimal places. Due to this rounding, the sum of cause-specific death rates within an age at times does not equal the all-cause death rate for the same age. We do not adjust for this artifact in the abridged data files given, which remain true to the original. Data files that we graduate to single ages are constrained to sum to the original abridged death counts within causes, and then reconstrained to sum between causes to the all-cause abridged death counts. The graduation/extrapolation procedure is described with more detail in a later section.

We change the data formatting in a few ways. First, data is reshaped to long format, yielding a single row for each population, year, sex, cause, and age. Datasets are available for each specific population, as well as in a single long file. We store datasets in `Rdata` binary files with no further rounding, and as `csv` files with rates rounded at the 6th decimal place.

2.1 Cause of death codes

Causes are identified using two-digit codes, where `"00"` is all-cause mortality, and `"01"`, ..., `"12"` are causes of death, given in the same order as the original published tables. Table 1 shows the correspondence of causes and codes used. Data files contain these codes rather than full cause-of-death names.

2.2 Other codes

Table 1: Cause codes used in the data files

Code	Name
00	All
01	Respiratory tuberculosis
02	Other infectious and parasitic diseases
03	Malignant and benign neoplasms
04	Cardiovascular diseases
05	Influenza, pneumonia, bronchitis
06	Diarrhea, gastritis, enteritis
07	Certain degenerative diseases
08	Complications of pregnancy
09	Certain diseases of infancy
10	Motor vehicle accidents
11	Other accidents and violence
12	All other and unknown causes

We include a column `Deaths`, which is calculated as $M(x) * P(x)$, except for all-cause mortality, which is taken directly from the original table. `Sex` is specified as 1 for males and 2 for females. Ages are given as lower age bound integer values in the `Age` column, obtaining values 0, 1, 5, 10,. . . , 85 in the abdriged files, and 0, 1, 2, . . . , 100 in the graduated files. Age intervals are given explicitly in the `AgeInterval`, taking values 1, 4, 5, 5, . . . , "+" for the original abridged data and 1, 1, 1,. . . for the graduated data.[1]

The orignal table titles are coded into new columns for `Country`, `Region`, and `Ethnicity` (where applicable) using codes adopted (and adapted) from the Human Lifetable Database. Country codes are given in Table 2

`Region` codes are only applicable in a few cases in the original data, as given in Table 3.

Race and ethnicity codes are only applicable in a few cases in the original data, as given in Table 4.

[1] We close out extrapolation at age 100 in the present version rather than treating age 100 as an open age group.

Table 2: Country codes used in the data files

Country	Code
Australia	AUS
Austria	AUT
Belgium	BEL
Bulgaria	BGR
Canada	CAN
Ceylon	LKA
Chile	CHL
Colombia	COL
Costa Rica	CRI
Czechoslovakia	CSK
Denmark	DNK
El Salvador	SVN
Finland	FIN
France	FRA
Germany	DEU
Greece	GRC
Guatemala	GTM
Hong Kong	HKG
Hungary	HUN
Iceland	ISL
Ireland	IRL
Israel	ISR
Italy	ITA
Japan	JPN
Malta and Gozo	MLT
Mauritius	MUS
Mexico	MEX
Netherlands	NLD
New Zealand	NZL
Norway	NOR
Panama	PAN
Philippines	PHL
Poland	POL
Portugal	PRT
Puerto Rico	PRI
South Africa	ZAF
Spain	ESP
Sweden	SWE
Switzerland	CHE
Taiwan	TWN
Trinidad & Tobago	TTO
United Kingdom	GBR
USA	USA
Venezuela	VEN
Yugoslavia	YUG

Table 3: Region codes used in the data files

Region	Code
England & Wales	FRG
former Federal Republic	ENW
former West Berlin	NIR
Northern Ireland	SCO
Scotland	GWB
Registration States	RS

Table 4: Race, ethnicity, and religion codes used in the data files

Ethnicity	Code
Non White	E092
White	E110
Coloured, Non White	E040

2.3 A glimpse at the data

The header and first six lines of the abridged data file is shown in Table 5. To repeat, the mortality rates in this file are true to the original, and these data have an open age group of 85+. Values are identical in the `Rdata` and `csv` files.

Table 5: A sample of the abridged data file.

Country	Region	Ethnicity	Year	Sex	Cause	Age	AgeInt	Population	Deaths	Mx
AUS			1911	1	00	0	1	60553	4750	0.078440
AUS			1911	1	00	1	4	213422	1386	0.006490
AUS			1911	1	00	5	5	235222	515	0.002190
AUS			1911	1	00	10	5	221100	357	0.001610
AUS			1911	1	00	15	5	232399	556	0.002390
AUS			1911	1	00	20	5	233779	852	0.003640

The header and first six lines of the graduated data file is shown in Table 6. To repeat, these data have been split into single ages and extrapolated to age 100. The `Rdata` files contain unrounded values, and the `csv` files contain values rounded to the 6th decimal place. The graduated data files do not contain `Population` or `Deaths` columns.

Table 6: A sample of the graduated data file

Country	Region	Ethnicity	Year	Sex	Cause	Age	AgeInt	Mx
AUS			1911	2	00	0	1	0.062330
AUS			1911	2	00	1	1	0.008907
AUS			1911	2	00	2	1	0.006535
AUS			1911	2	00	3	1	0.004819
AUS			1911	2	00	4	1	0.003604
AUS			1911	2	00	5	1	0.002763

3 Quality of original data

The original authors discuss data collection criteria, coding practices, comparability, data quality, and methods employed at length, as so we only summarize them briefly here. Death counts were taken as published by statistical offices, and these were not modified in any way. Lifetables were typically published to coincide with census-years, but alternative sources or methods were used for population denominators when censuses were not available for the reference year.

The twelve causes of death were originally selected based on a variety of considerations, which we here paraphrase. Codes were designed with etiological considerations in mind, to facilitate separation between internal and external (organic and inorganic) causes of death. Other practical constraints included i) that the causes needed to be separable from existing cause classifications, ii) easily confounded diseases were grouped, iii) ill-definied and unknown causes needed to be separately identified, and iv) the number of causes was also reduced via grouping in order to facilitate processing at the time. The reference table used to aggregate finer causes of death given in each ICD revision into the 12 codes used in the book is given in Appendix Table 1.

The original authors conducted diagnostics of census completeness and reliability using indirect methods, with summary results reproduced in Appendix Tables 7 and 8. Six countries are coded as as having underregistration in the census (Sri Lanka (Ceylon), Colombia, Greece, Panama, Philippines and Venezuela) but the majority show acceptable or good quality denominators.

The original authors also checked causes of death data and found some problems. In cases of coding inconsistency between countries, codes were adjusted to conform as well as possible with the given set of 12 causes. Nevertheless some data quality issues are described in the original book, such as underregistration of motor vehicle mortality in Latin America. We have carried out external validity checks by comparing all-cause mortality rates to those from the Human Mortality Database (Human Mortality Database) where comparable estimates were available, and with the Human Life-Table Database in some other cases. These comparisons were in nearly all instances very close. A detailed country-by country comparison is available in a separate data quality report.

4 Graduation of mortality rates to single ages

We graduate the orginal abridged data to single-age cause-specific mortality rates using the penalized composite link model of Rizzi, Gampe, and Eilers (Rizzi et al. 2015). This method allows for a smooth single age pattern, and it ensures that single age counts derived from each age group sum to the total from the original age group. Since infant mortality is already given in single ages, only ages 2 to 85 are graduated. Age 85 is given an interval of 15 years for purposes of graduation, such that each cause of death is closed out by age 100. We graduate in two states, first splitting exposures to single ages, and then using the derived single-age exposures as an offset in graduating the abridged death counts. We use a smoothing parameter λ=5.75 as suggested by Rizzi et al. (2015) for both exposures and death counts. Further details can be found in the `R` code, which will be made available on the main project website.

5 Conclusions

This digitil collection makes available an important database that for so long has been close at hand but inaccessible for most researchers. This database offers a deep and wide reach into the history of cause-of-death changes. While esimates are not free from error, results derived from these rates may still lead to new broad insights in population history. While we can confirm that all-cause estimates in this database are rather close to other high quality estimates, we cannot make such external comparisons for some populations available in this database, and we have not assessed the quality of registration or coding for specific causes of death. We therefore recommend to generate summary indices where possible, and to carry out individual cause-of-death diagnostics as required by particular research aims. Graduated rates are not of any higher quality than the original abridged rates, but are given simply for ease of use in common applications. Single-age rates beyond age 85 are to be used with caution. Future changes to this database will be limited to the graduation method, documentation, and dissemination, but the original abridged data will in all cases be left in tact as provided here.

References

Human Mortality Database. University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded February 15th, 2015).

Samuel H Preston, Nathan Keyfitz, and Robert Schoen. *Cause of Death: Life Tables for National Populations.* Seminar Press, 1972.

Silvia Rizzi, Jutta Gampe, and Paul H. C. Eilers. Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology Advance Access*, 182:138 – 147, June 16 2015. doi: doi: 10.1093/aje/kwv020.

Appendix A Diagnostics reproduced from the original authors.

Figure 1: Compostion of Cause-of-Death Categories

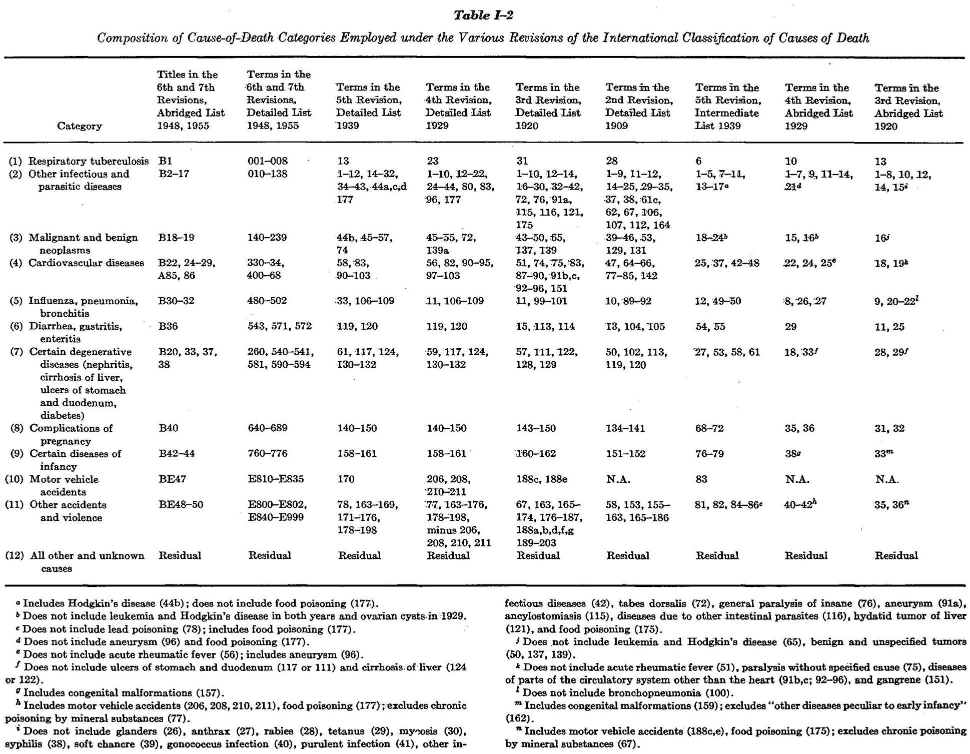

Table I–2

Composition of Cause-of-Death Categories Employed under the Various Revisions of the International Classification of Causes of Death

Category	Titles in the 6th and 7th Revisions, Abridged List 1948, 1955	Terms in the 6th and 7th Revisions, Detailed List 1948, 1955	Terms in the 5th Revision, Detailed List 1939	Terms in the 4th Revision, Detailed List 1929	Terms in the 3rd Revision, Detailed List 1920	Terms in the 2nd Revision, Detailed List 1909	Terms in the 5th Revision, Intermediate List 1939	Terms in the 4th Revision, Abridged List 1929	Terms in the 3rd Revision, Abridged List 1920
(1) Respiratory tuberculosis	B1	001–008	13	23	31	28	6	10	13
(2) Other infectious and parasitic diseases	B2–17	010–138	1–12, 14–32, 34–43, 44a,c,d 177	1–10, 12–22, 24–44, 80, 83, 96, 177	1–10, 12–14, 16–30, 32–42, 72, 76, 91a, 115, 116, 121, 175	1–9, 11–12, 14–25, 29–35, 37, 38, 61c, 62, 67, 106, 107, 112, 164	1–5, 7–11, 13–17[a]	1–7, 9, 11–14, 21[d]	1–8, 10, 12, 14, 15[i]
(3) Malignant and benign neoplasms	B18–19	140–239	44b, 45–57, 74	45–55, 72, 139a	43–50, 65, 137, 139	39–46, 53, 129, 131	18–24[b]	15, 16[b]	16[j]
(4) Cardiovascular diseases	B22, 24–29, A85, 86	330–34, 400–68	58, 83, 90–103	56, 82, 90–95, 97–103	51, 74, 75, 83, 87–90, 91b,c, 92–96, 151	47, 64–66, 77–85, 142	25, 37, 42–48	22, 24, 25[e]	18, 19[k]
(5) Influenza, pneumonia, bronchitis	B30–32	480–502	33, 106–109	11, 106–109	11, 99–101	10, 89–92	12, 49–50	8, 26, 27	9, 20–22[l]
(6) Diarrhea, gastritis, enteritis	B36	543, 571, 572	119, 120	119, 120	15, 113, 114	13, 104, 105	54, 55	29	11, 25
(7) Certain degenerative diseases (nephritis, cirrhosis of liver, ulcers of stomach and duodenum, diabetes)	B20, 33, 37, 38	260, 540–541, 581, 590–594	61, 117, 124, 130–132	59, 117, 124, 130–132	57, 111, 122, 128, 129	50, 102, 113, 119, 120	27, 53, 58, 61	18, 33[f]	28, 29[f]
(8) Complications of pregnancy	B40	640–689	140–150	140–150	143–150	134–141	68–72	35, 36	31, 32
(9) Certain diseases of infancy	B42–44	760–776	158–161	158–161	160–162	151–152	76–79	38[g]	33[m]
(10) Motor vehicle accidents	BE47	E810–E835	170	206, 208, 210–211	188c, 188e	N.A.	83	N.A.	N.A.
(11) Other accidents and violence	BE48–50	E800–E802, E840–E999	78, 163–169, 171–176, 178–198	77, 163–176, 178–198, minus 206, 208, 210, 211	67, 163, 165–174, 176–187, 188a,b,d,f,g 189–203	58, 153, 155–163, 165–186	81, 82, 84–86[c]	40–42[h]	35, 36[n]
(12) All other and unknown causes	Residual	Residual	Residual	Residual	Residual	Residual	Residual	Residual	Residual

[a] Includes Hodgkin's disease (44b); does not include food poisoning (177).

[b] Does not include leukemia and Hodgkin's disease in both years and ovarian cysts in 1929.

[c] Does not include lead poisoning (78); includes food poisoning (177).

[d] Does not include aneurysm (96) and food poisoning (177).

[e] Does not include acute rheumatic fever (56); includes aneurysm (96).

[f] Does not include ulcers of stomach and duodenum (117 or 111) and cirrhosis of liver (124 or 122).

[g] Includes congenital malformations (157).

[h] Includes motor vehicle accidents (206, 208, 210, 211), food poisoning (177); excludes chronic poisoning by mineral substances (77).

[i] Does not include glanders (26), anthrax (27), rabies (28), tetanus (29), mycosis (30), syphilis (38), soft chancre (39), gonococcus infection (40), purulent infection (41), other infectious diseases (42), tabes dorsalis (72), general paralysis of insane (76), aneurysm (91a), ancylostomiasis (115), diseases due to other intestinal parasites (116), hydatid tumor of liver (121), and food poisoning (175).

[j] Does not include leukemia and Hodgkin's disease (65), benign and unspecified tumors (50, 137, 139).

[k] Does not include acute rheumatic fever (51), paralysis without specified cause (75), diseases of parts of the circulatory system other than the heart (91b,c; 92–96), and gangrene (151).

[l] Does not include bronchopneumonia (100).

[m] Includes congenital malformations (159); excludes "other diseases peculiar to early infancy" (162).

[n] Includes motor vehicle accidents (188c,e), food poisoning (175); excludes chronic poisoning by mineral substances (67).

(Preston et al. 1972)

Table 7: Accuracy and Completeness of Systems of Census and Death Registration 1

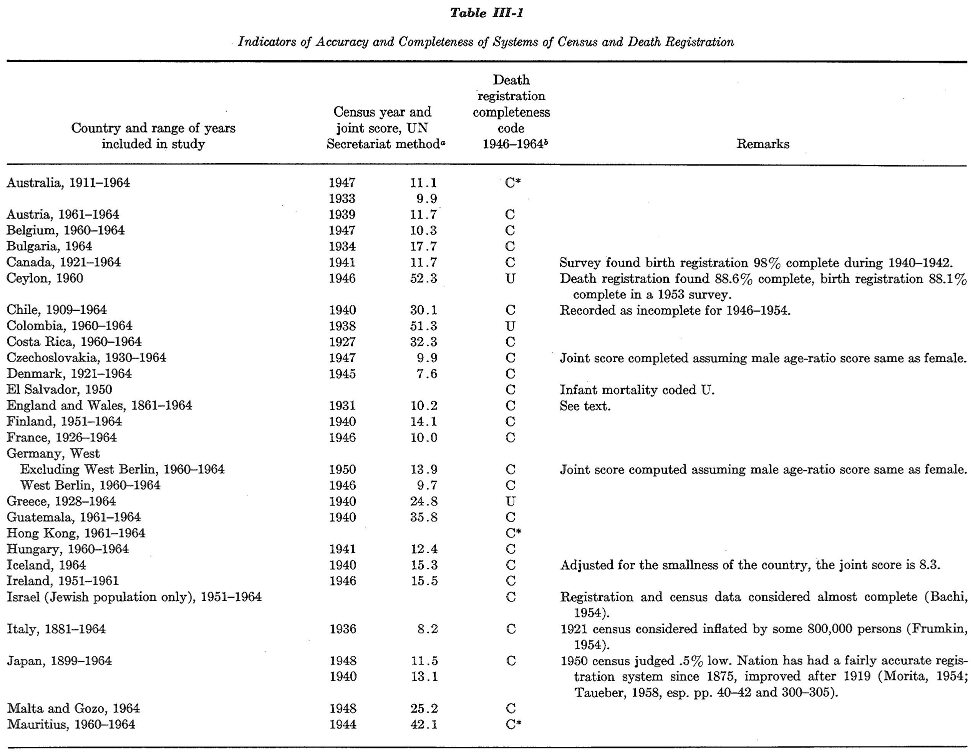

Table III-1

Indicators of Accuracy and Completeness of Systems of Census and Death Registration

Country and range of years included in study	Census year and joint score, UN Secretariat method[a]		Death registration completeness code 1946–1964[b]	Remarks
Australia, 1911–1964	1947	11.1	C*	
	1933	9.9		
Austria, 1961–1964	1939	11.7	C	
Belgium, 1960–1964	1947	10.3	C	
Bulgaria, 1964	1934	17.7	C	
Canada, 1921–1964	1941	11.7	C	Survey found birth registration 98% complete during 1940–1942.
Ceylon, 1960	1946	52.3	U	Death registration found 88.6% complete, birth registration 88.1% complete in a 1953 survey.
Chile, 1909–1964	1940	30.1	C	Recorded as incomplete for 1946–1954.
Colombia, 1960–1964	1938	51.3	U	
Costa Rica, 1960–1964	1927	32.3	C	
Czechoslovakia, 1930–1964	1947	9.9	C	Joint score completed assuming male age-ratio score same as female.
Denmark, 1921–1964	1945	7.6	C	
El Salvador, 1950			C	Infant mortality coded U.
England and Wales, 1861–1964	1931	10.2	C	See text.
Finland, 1951–1964	1940	14.1	C	
France, 1926–1964	1946	10.0	C	
Germany, West				
Excluding West Berlin, 1960–1964	1950	13.9	C	Joint score computed assuming male age-ratio score same as female.
West Berlin, 1960–1964	1946	9.7	C	
Greece, 1928–1964	1940	24.8	U	
Guatemala, 1961–1964	1940	35.8	C	
Hong Kong, 1961–1964			C*	
Hungary, 1960–1964	1941	12.4	C	
Iceland, 1964	1940	15.3	C	Adjusted for the smallness of the country, the joint score is 8.3.
Ireland, 1951–1961	1946	15.5	C	
Israel (Jewish population only), 1951–1964			C	Registration and census data considered almost complete (Bachi, 1954).
Italy, 1881–1964	1936	8.2	C	1921 census considered inflated by some 800,000 persons (Frumkin, 1954).
Japan, 1899–1964	1948	11.5	C	1950 census judged .5% low. Nation has had a fairly accurate registration system since 1875, improved after 1919 (Morita, 1954; Taueber, 1958, esp. pp. 40–42 and 300–305).
	1940	13.1		
Malta and Gozo, 1964	1948	25.2	C	
Mauritius, 1960–1964	1944	42.1	C*	

(Preston et al. 1972)

Table 8: Accuracy and Completeness of Systems of Census and Death Registration 2

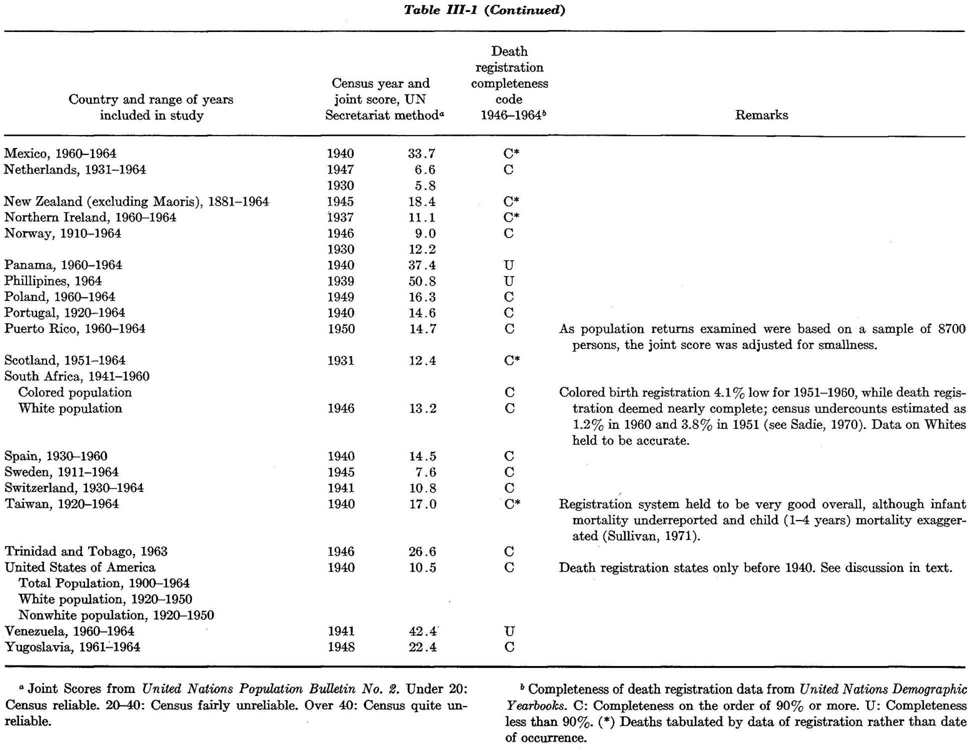

Table III-1 (Continued)

Country and range of years included in study	Census year and joint score, UN Secretariat method[a]		Death registration completeness code 1946–1964[b]	Remarks
Mexico, 1960–1964	1940	33.7	C*	
Netherlands, 1931–1964	1947	6.6	C	
	1930	5.8		
New Zealand (excluding Maoris), 1881–1964	1945	18.4	C*	
Northern Ireland, 1960–1964	1937	11.1	C*	
Norway, 1910–1964	1946	9.0	C	
	1930	12.2		
Panama, 1960–1964	1940	37.4	U	
Phillipines, 1964	1939	50.8	U	
Poland, 1960–1964	1949	16.3	C	
Portugal, 1920–1964	1940	14.6	C	
Puerto Rico, 1960–1964	1950	14.7	C	As population returns examined were based on a sample of 8700 persons, the joint score was adjusted for smallness.
Scotland, 1951–1964	1931	12.4	C*	
South Africa, 1941–1960				
Colored population			C	Colored birth registration 4.1% low for 1951–1960, while death registration deemed nearly complete; census undercounts estimated as 1.2% in 1960 and 3.8% in 1951 (see Sadie, 1970). Data on Whites held to be accurate.
White population	1946	13.2	C	
Spain, 1930–1960	1940	14.5	C	
Sweden, 1911–1964	1945	7.6	C	
Switzerland, 1930–1964	1941	10.8	C	
Taiwan, 1920–1964	1940	17.0	C*	Registration system held to be very good overall, although infant mortality underreported and child (1–4 years) mortality exaggerated (Sullivan, 1971).
Trinidad and Tobago, 1963	1946	26.6	C	
United States of America	1940	10.5	C	Death registration states only before 1940. See discussion in text.
Total Population, 1900–1964				
White population, 1920–1950				
Nonwhite population, 1920–1950				
Venezuela, 1960–1964	1941	42.4	U	
Yugoslavia, 1961–1964	1948	22.4	C	

[a] Joint Scores from *United Nations Population Bulletin No. 2.* Under 20: Census reliable. 20–40: Census fairly unreliable. Over 40: Census quite unreliable.

[b] Completeness of death registration data from *United Nations Demographic Yearbooks.* C: Completeness on the order of 90% or more. U: Completeness less than 90%. (*) Deaths tabulated by data of registration rather than date of occurrence.

(Preston et al. 1972)