Chapter 3
Data and Software

3.1 Data

3.1.1 Human Mortality Database

Most of our analyses are based on data from the Human Mortality Database ("HMD", 2017), which can be freely accessed after registration at http://www.mortality.org. The database is a collaborative project of research teams from the Department of Demography at the University of California, Berkeley (USA) and the Max Planck Institute for Demographic Research in Rostock (Germany). It contains *aggregate* mortality statistics such as death counts, population estimates, exposure to risk estimates, life tables as well as some other statistics of more than 35 countries (see Table 3.1). Further distinctions into sub populations are possible for some countries such as Germany (East and West Germany), the United Kingdom (England and Wales, Northern Ireland, Scotland) or New Zealand (Maori, Non-Maori). The database has its focus on highly developed countries.

Since its launch in 2002, the HMD has become the gold standard for the aggregate level (demographic) analysis of mortality. Apart from the diligent collection of data, its widespread adoption can mainly be attributed to two reasons: (1) Rigorous quality checks are conducted before new data are added to the database. (2) The biggest asset of the HMD is that it does not simply publish processed data. Instead, the HMD estimates life tables and other statistics itself using raw data, applying the same set of methods. Thus, any differences over time or across region can not be attributed to different methodologies, for instance, how the life table was closed (HMD 2007).

As some life tables in the HMD are smoothed at ages 80 and higher, we did not rely on life tables estimates at all but used exclusively the death counts and the corresponding exposures from the HMD on a 1-calendar-year by 1-age-year grid to estimate death rates. Most of our analyses deal with mortality developments

R. Rau et al., *Visualizing Mortality Dynamics in the Lexis Diagram*,
The Springer Series on Demographic Methods and Population Analysis 44,
DOI 10.1007/978-3-319-64820-0_3

Table 3.1 Countries covered in the Human Mortality Database and data coverage after 1950 on January 10th, 2017, when the most recent update of data was conducted for the present monograph

Country	Deaths	Data Coverage
Australia	6,956,698	
Austria	5,575,409	
Belarus	5,739,008	
Belgium	7,216,189	
Bulgaria	5,663,709	
Canada	11,056,710	
Chile	1,121,837	
Czech Republic	7,366,321	
Denmark	3,385,967	
Estonia	933,627	
Finland	3,025,080	
France	34,953,421	
Germany	20,660,306	
Germany−East	11,839,606	
Germany−West	40,080,602	
Greece	3,314,273	
Hungary	8,264,444	
Iceland	100,400	
Ireland	2,107,181	
Israel	1,132,519	
Italy	33,542,107	
Japan	54,561,300	
Latvia	1,667,906	
Lithuania	1,953,998	
Luxembourg	217,818	
Netherlands	7,242,595	
New Zealand	1,610,320	
New Zealand: Maori	94,938	
New Zealand: Non−Maori	1,368,240	
Norway	2,563,597	
Poland	18,949,583	
Portugal	6,296,406	
Russia	90,757,552	
Slovakia	2,988,888	
Slovenia	610,127	
Spain	20,721,634	
Sweden	5,590,479	
Switzerland	3,773,553	
Taiwan	4,961,642	
UK	40,029,614	
UK, England & Wales	35,120,479	
UK, Northern Ireland	995,217	
UK, Scotland	3,913,918	
Ukraine	31,897,144	
USA	133,314,965	
		1950 1960 1970 1980 1990 2000 2010

since 1950. We selected this threshold year because of the availability of more data compared to earlier time periods. Furthermore, it also marks the beginning of a new era: Most gains in life expectancy are nowadays due to survival improvements among the elderly (Christensen et al. 2009), a development, which was virtually non-existent before the middle of the twentieth century. Kannisto (1994), for instance, estimated that the onset of sustained decline in old-age mortality occurred for women in Switzerland, Belgium and Sweden in 1956.

As shown in Table 3.1 total deaths range from barely 100,000 (Iceland) to more than 130 million in the United States. We analyzed all countries; the only exceptions are Chile and the Maori population of New Zealand due to problematic data quality (Jdanov et al. 2008) and the low number of years covered (Chile). Nevertheless, we did not include those figures for all countries and both sexes as it would have resulted in a monograph consisting of hundreds of additional pages. We typically restricted ourselves, instead, to a few examples that feature interesting characteristics.

3.1.2 Cause-Specific Death Counts in the United States

The National Center for Health Statistics of the United States provides a unique collection: Individual death counts by sex, age at death, year of death, cause of death, and many more characteristics can be freely downloaded from its web page. The data are available since 1968 in annual files. Additionally, the website of the National Bureau of Economic Research (NBER) provides data since 1959, which we used in our analyses. The last year in our analysis is 2014. With the exception of 1972, when only a 50% sample was taken, each file contains all deaths in the United States. In the analysis by cause of death in later chapters of this volume, we simply multiplied the number of deaths for a given age, sex, and cause in the year 1972 by a factor of 2.

Causes of death are coded by the so-called "International Classification of Diseases" (ICD). Since its introduction in the late nineteenth century, the system has been revised at irregular intervals (Meslé 2006). The tenth revision is currently used. During the first years of our analysis, ICD-7 was used. ICD-8 was in effect in the United States between 1968 and 1978, followed by ICD-9 from 1979 until 1998.

Obtaining consistent time series of causes of death across ICD revisions requires meticulous work and care (e.g., Meslé and Vallin 1996; Pechholdová 2009). We therefore decided to use only very broad categories for causes of death and followed primarily the coding of Janssen et al. (2003) and of Meslé and Vallin (2006a). Both papers include an appendix with detailed ICD codes across the four revisions required in our analysis.

Table 3.2 is split into two halves. The upper panel provides the ICD codes we used to extract the causes of death, whereas the lower panel lists the number of deaths in absolute and relative terms for the selected causes by sex.

Our database consists of more than 118 million deaths. Although we have selected very few causes, they account for about three quarters of all deaths (Category 13 "Other" is 23.75%). A bit more than 44% of all deaths classified as originating from circulatory diseases. In that category, heart diseases are about one third of all deaths for women and men alike. The almost 10 million deaths from cerebrovascular diseases between 1959 and 2014 represent about eight percent of all deaths. The most common cerebrovascular disease is stroke. Malignant neoplasms ("cancer") are the second largest chapter in the ICD. Regardless of sex of the decedent, about one in every fifth death belongs to that category. We

Table 3.2 ICD codes and counts (absolute and relative) for females, males, and both sexes combined selected causes of death, 1959–2014

	Cause	ICD codes			
		ICD-7	ICD-8	ICD-9	ICD-10
Nr.	Years in use:	1959–1967	1968–1978	1979–1998	1999–2014
(1)	All causes	—	—	—	—
(2)	Circulatory dis.	300–334, 400–468	390–458	390–459	I00–I99
(3)	Heart	400–447	390–429	390–429	I00–I52
(4)	Cerebrovasc.	300–334,	430–434, 436–438	430–434, 436–438	I60–I69
(5)	Other	All (2) not in (3) or (4)			
(6)	Cancers	140–239	140–239	140–239	C00–D48
(7)	Breast	170	174	174, 175	C50
(8)	Lung	162, 163	162	162	C33, C34
(9)	Colorectum	153, 154	153, 154	153, 154	C18–C21
(10)	Other	All (6) not in (7), (8), or (9)			
(11)	Resp. diseases	470–527	460–519	460–519	J00–J99
(12)	Motor vehicle acc.	E810–E825	E810–E819	E810–E819	V00–V89
(13)	Other	All (1) not in (2)–(12)			

		Number of cases					
		Total		Female		Male	
Nr.	Cause	Counts	%	Counts	%	Counts	%
(1)	All causes	118,678,283	(100.00)	56,432,184	(100.00)	62,246,099	(100.00)
(2)	Circulatory dis.	52,668,448	(44.38)	25,985,900	(46.05)	26,682,548	(42.87)
(3)	Heart	40,342,012	(33.99)	19,072,073	(33.80)	21,269,939	(34.17)
(4)	Cerebrovasc.	9,381,071	(7.90)	5,430,076	(9.62)	3,950,995	(6.35)
(5)	Other	2,945,365	(2.48)	1,483,751	(2.63)	1,461,614	(2.35)
(6)	Cancers	25,722,893	(21.67)	12,096,049	(21.43)	13,626,844	(21.89)
(7)	Breast	2,067,878	(1.74)	2,050,192	(3.63)	17,686	(0.03)
(8)	Lung	6,393,007	(5.39)	2,260,023	(4.00)	4,132,984	(6.64)
(9)	Colorectum	2,884,519	(2.43)	1,458,772	(2.59)	1,425,747	(2.29)
(10)	Other	14,377,489	(12.11)	6,327,062	(11.21)	8,050,427	(12.93)
(11)	Resp. diseases	9,566,798	(8.06)	4,457,141	(7.90)	5,109,657	(8.21)
(12)	Motor vehicle acc.	2,538,449	(2.14)	742,599	(1.32)	1,795,850	(2.89)
(13)	Other	28,181,695	(23.75)	13,150,495	(23.30)	15,031,200	(24.15)

selected three prominent cancer sites: Breast, lung and colorectum. Please note that while there are many more deaths from breast cancer for women, also more than 17,000 men died from it during the 56 years of our observation period. Respiratory diseases are with approximately 8% of all deaths slightly more common than cerebrovascular diseases. Although it is not a major cause of death (2%), we also

included information about motor vehicle accidents since it turned out to be an interesting case study for seasonality in deaths, which we analyze in Chap. 9.

3.1.3 SEER Cancer Register Data 1973–2011

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute of the United States allows researchers access to longitudinal data on the individual level about the incidence of cancer and includes also information about the survival of patients. The data coverage—the SEER data start in 1973—and the large size of data, combined with the ease of access, make the SEER data an ideal instrument for the analysis of cancer survival by age over calendar time. We were using data that were released in April 2014 with a follow-up cutoff date of December 31, 2011 (Surveillance, Epidemiology, and End Results (SEER) Program 2014). The SEER data do *not* cover all cancer diagnoses of the United States. It is a collection of data from several registries. With the exception of Seattle (Puget Sound) and Metropolitan Atlanta that started in 1974 and 1975, respectively, we only used registers that covered the whole time span from 1973 until the end of 2011. Although we use less data than we could have, we thought that a heterogeneous set of registers would have induced problems for the analysis over time. The registers included in our analysis were: San Francisco-Oakland SMSA, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Utah as well as Seattle and Metropolitan Atlanta.

In our analysis of cancer survival in Chap. 10, starting on page 123, we selected five cancer sites: Breast cancer; cancer of the lung and bronchus; cancer of the colon, rectum, and anus; pancreatic cancer; prostate cancer. As shown in Table 3.3, those five cancer sites constitute about 55% of all cancer diagnoses for women as well as for men out of the 4.5 million cases recorded during our observation period. The largest categories are by far breast cancer for women (30.44%) and prostate cancer for men (25.79%). The absolute and relative frequencies of the other cancer sites as well as their respective ICD codes can be inspected from Table 3.3. While ICD-8 was in use at the beginning of the observation period in 1973 and cancer cases are typically coded by the ICD-O standard, all ICD codes were converted to ICD-10 by SEER.

3.2 Software

All analyses have been conducted and all figures have been produced using R (Version 3.2.3), a free software environment for statistical computing and graphics (R Development Core Team 2015). The surface maps were created by the `image()` function and contour lines were added with the `contour()` function. To facilitate the creation of surface maps of rates of mortality improvement for other researchers,

Table 3.3 ICD-10 codes and incidence counts (absolute and relative) by cancer site of females, males, and both sexes combined in the SEER Data, 1973–2011

Cancer site		ICD-10 Code	Incidence Total Counts	in %
(1)	All	C00–D48	4,524,099	(100.00)
(2)	Breast	C50	713,376	(15.77)
(3)	Bronchus and lung	C34	557,901	(12.33)
(4)	Colon, rectum, and anus	C18-C21	520,456	(11.50)
(5)	Pancreas	C25	103,152	(2.28)
(6)	Prostate	C61	566,311	(12.52)
(7)	Rest	All (1) not in (2)–(6)	2,062,903	(45.60)

Cancer site		Incidence Female Counts	in %	Male Counts	in %
(1)	All	2,328,116	(100.00)	2,195,983	(100.00)
(2)	Breast	708,696	(30.44)	4,680	(0.21)
(3)	Bronchus and lung	224,927	(9.66)	332,974	(15.16)
(4)	Colon, rectum, and anus	257,406	(11.06)	263,050	(11.98)
(5)	Pancreas	51,712	(2.22)	51,440	(2.34)
(6)	Prostate	N/A	(N/A)	566,311	(25.79)
(7)	Rest	1,085,375	(46.62)	977,528	(44.51)

an R package called `ROMIplot` has been created and uploaded to CRAN, the general archive of R packages. Installation and usage of this package are explained in Appendix "Software: R package `ROMIplot`" (p. 161).