3

Health Outcome #1: Child Survival

Child mortality is a commonly used measure of average population health (International Monetary Fund et al. 2000; UNICEF 2001) and has been used in studies of the gaps in health outcomes between the poor and the better-off (Gwatkin et al. 2000; Wagstaff 2000). The infant mortality rate (IMR) is the number of deaths occurring in the first year of life per 1,000 live births and measures the probability of a child dying before the child's first birthday. The under-five mortality rate (U5MR) measures the probability of death before a child's fifth birthday. Although the IMR and U5MR can be estimated from vital registration statistics and demographic surveillance systems, these are not widely used in the developing world. In instances in which vital registration systems do exist, their comprehensiveness and reliability are often doubted. Furthermore, even where they exist, it is uncommon for any socioeconomic information to be recorded, making the analysis of socioeconomic inequalities in mortality impossible.

The alternative source of data on child mortality is a household survey. Estimating mortality rates from survey data involves the use of fertility histories. These are constructed from responses to questions posed to women of fertile age about births and deaths of children born to them. A complete fertility history uncovers the dates of birth, and if applicable the deaths, of all children born to the interviewed woman. An incomplete fertility history uncovers only the number of children born to the interviewed woman and the number still alive (or equivalently the number who have died).

Complete and incomplete fertility histories call for two different methods of mortality estimation, the direct method and the indirect method, respectively. Each method has advantages and disadvantages. The complete fertility history places greater informational demands on the survey and the interviewed woman, but generates more information and permits the estimation of standard errors for the mortality estimates. The incomplete history is less demanding in regard to information, but requires that the survey data be supplemented with data from a model life table, and it is not possible to compute standard errors for the mortality estimates.

This chapter describes how to compute mortality rates from household survey data by the direct and indirect methods using two statistical packages. It also explains how survey data can be used to undertake disaggregated mortality estimation, for example, across socioeconomic groups.

Complete fertility history and direct mortality estimation

The direct method of estimating child mortality involves taking the data from the complete fertility history and estimating a life table. This section outlines the steps involved and sets out the Stata code using a worked example from the 1998 Vietnam Living Standards Survey (VLSS) (Wagstaff and Nguyen 2003).

Preparing the data: example from the Living Standards Measurement Study

The complete fertility history allows a child-level data set to be assembled containing data for each child on the date of the mother's interview, the date of birth of the child, a binary variable indicating the child's status (alive or dead), and the survival time. Mortality estimation is based on deaths over a specific period of time—usually the 5 or 10 years before the survey date. It is useful therefore to have a variable indicating how old the child was at the interview date in the case of a still-living child and, in the case of a child who had died, how old the child would have been at the interview date if the child had still been alive at that time. This variable, called `hypage` below, allows one to easily select only those cases in which the child was born in the past 5 (or 10) years.

The first step is to generate the interview date—or more precisely the date the fertility history was collected. In a typical living standards measurement study (LSMS) survey (for Demographic and Health Surveys [DHS] see below) this date is recorded in the file preceding the household roster. In the first line of code below `date2` is the variable recording the date the fertility history was collected. In this particular example it was recorded as a numeric variable in the form *ddmmyy*, where *dd* is the day, *mm* the month, and *yy* the last two digits of the year. The month, *mm*, is entered as two digits (e.g., April was entered as 04 rather than 4), but the day, *dd*, had no trailing zero (e.g., 1 April is entered as 104 rather than 0104). The first few commands generate three new numeric variables corresponding to the day, the month, and the year (in full—i.e., as 1997, rather than 97) when the fertility history was collected. The `mdy` function puts this into the date variable `intrvdate2`. The data are then sorted by household ID (`househol`), in anticipation of a merge with the fertility history data, and saved.

```
gen intrvdate = date2
tostring intrvdate
gen str2 year = substr(intrvdate,-2,.)
gen str2 mnth = substr(intrvdate,-4,.)
destring year , replace
replace year = 1900+year
destring mnth , replace
gen day = int(date2/10000)
gen intrvdate2 = mdy(mnth,day,year)
format %d intrvdate2
sort househol
save filename, replace
```

The next step is to generate each child's date of birth (`dob`) in the fertility history data file. In the commands below, `s8aq05y` denotes the last two digits of the year of birth of the child, `s8aq05d` is the day of the month the child was born, and `s8aq05m` is the month. In all fertility histories, there is the problem of what to do with cases in which some information in the date of birth is missing. One option is to drop such cases, but this reduces the sample size. Moreover, because it is likely that these cases are more often than not children who have died, dropping them will bias downward the estimated mortality rate. An alternative (the procedure adopted below) is to drop cases in which the year of birth is missing, but to replace missing months by 6 and missing days by 15. The `mdy` command creates a date variable corresponding to the `dob`.

```
gen yob = 1900+s8aq05y
gen daybrth = s8aq05d
replace daybrth = 15 if s8aq05d==.
gen mob = s8aq05m
replace mob = 6 if s8aq05m==.
gen dob = mdy(mob,daybrth,yob)
format %d dob
sort househol
```

Next the fertility history data (the open file, or master file) and the interview data (the using data set) are merged using `merge`, and variables that measure the child's hypothetical age in days (`hypagedays`) and years (`hypageyrs`) are computed.

```
merge househol using filename
gen hypagedays = intrvdate2 - dob
gen hypageyrs = hypagedays/365
```

The next step is to generate a status variable, `dead`, taking a value of 1 if the child has died and zero otherwise. In the commands below `s8aq07` is the relevant question (Is child *x* still living with your household?), 1 being "yes", 2 being "no, living in another place," and 3 being "no, died."

```
gen dead =(s8aq07==3)
```

Finally, we need to generate a variable measuring the survival time. This is equal to the time elapsed since birth in the case of children who are still alive and equal to time between birth and death in the case of children who have died. In the 1998 VLSS, `s8aq09t` measures the survival time of children who have died, and `s8aq09u` indicates the units in which the survival time is measured (minutes, hours, days, weeks, months, quarters, half years, or years). Here we have computed two variables—`timedays` and `timeyears`.

```
gen timedays = hypagedays
replace timedays = s8aq09t/1440 if dead==1 & s8aq09u==1
replace timedays = s8aq09t/24 if dead==1 & s8aq09u==2
replace timedays = s8aq09t if dead==1 & s8aq09u==3
replace timedays = s8aq09t*7 if dead==1 & s8aq09u==4
replace timedays = s8aq09t*30.42 if dead==1 & s8aq09u==5
replace timedays = s8aq09t*91.25 if dead==1 & s8aq09u==6
replace timedays = s8aq09t*182.5 if dead==1 & s8aq09u==7
replace timedays = s8aq09t*365 if dead==1 & s8aq09u==8
gen timeyears = timedays / 365
```

Preparing the data: Demographic and Health Survey

The DHS uses a century month code (`CMC`) for some of its date variables. A `CMC` is the number of the month since the start of the century. For example, January 1900 is `CMC 1;` January 1901 is `CMC 13`. The variables needed for direct mortality estimation in DHS are

```
V008: date of interview (CMC)
B3: date of birth (CMC)
B7: age at death (month imputed)
B5: whether the child is still alive
```

The age at interview variable, `hypage`, can be calculated for children alive and children dead in Stata as follows:

```
gen hypage=(v008-b3)/12
```

Then, we can generate the surviving time (in years), `timeyears`, for each child in the survey and an indicator of whether the child is `dead`

```
gen timeyears=.
replace timeyears=hypage
replace timeyears=b7/12 if b5==0
gen dead=(b5==0)
```

Computing mortality rates and standard errors

The IMR and U5MR are computed using a life table, produced using the command `ltable`. The command below selects only those children born in the previous 10 years, including those born exactly 10 years ago. Stata allows one to specify the interval width, which can vary through the life table.[1] In the case below, a fixed half-yearly interval is used.

```
ltable timeyears dead if hypageyrs <=10 , int(.5) gr
```

Stata produces a life table along the lines of table 3.1. The lack of decimals in the intervals makes interpretation somewhat difficult—the first row refers to the first half-year of life, the second row to the second half-year of life, and so on. There were 5,316 children born during the previous 10 years, of whom 114 died during the first six months of life, and 194 were "lost" or censored—that is, they were born within six months of the interview date and were therefore not fully exposed to the risk of death. The assumption made in the life table is that these 194 children were exposed for only half of the interval—in this case three months rather than six. The total number of children exposed during the first six months is thus 5,316 less half of 194, or 5,219. The survival rate for the first six months is therefore (5,316 – 114) divided by 5,219, or 0.9782. The survival rate for each of the subsequent half-years is computed in the same way, and from these the cumulative survival function (labeled simply "survival" in table 3.1) is formed. The IMR is the complement of the cumulative survival function at the end of the first year—that is, 1 – 0.9752, or equivalently 24.8 per 1,000 live births. The U5MR is equal to 1 – 0.9642, or 35.8

---

[1] For example, one could divide the first year into months using `int(0.08333, 0.1666, …, 0.9167, 1.5, 2, 2.5, …, 10)`.

Table 3.1 *Life Table, Vietnam, 1988–98*

Interval		*Total*	*Deaths*	*Lost*	*Survival*	*Error*	*[95% Conf. Int.]*	
0	1	5316	114	194	0.9782	0.0020	0.9738	0.9818
1	1	5008	15	161	0.9752	0.0022	0.9706	0.9791
1	2	4832	12	176	0.9727	0.0023	0.9679	0.9768
2	2	4644	4	190	0.9719	0.0023	0.9670	0.9760
2	3	4450	15	198	0.9685	0.0025	0.9633	0.9730
3	3	4237	0	188	0.9685	0.0025	0.9633	0.9730
3	4	4049	13	227	0.9653	0.0026	0.9598	0.9701
4	4	3809	0	227	0.9653	0.0026	0.9598	0.9701
4	5	3582	4	231	0.9642	0.0027	0.9586	0.9690
5	5	3347	0	284	0.9642	0.0027	0.9586	0.9690
5	6	3063	7	314	0.9619	0.0028	0.9560	0.9670
6	6	2742	0	262	0.9619	0.0028	0.9560	0.9670
6	7	2480	3	263	0.9606	0.0029	0.9546	0.9659
7	7	2214	0	288	0.9606	0.0029	0.9546	0.9659
7	8	1926	3	297	0.9590	0.0030	0.9527	0.9645
8	8	1626	0	323	0.9590	0.0030	0.9527	0.9645
8	9	1303	0	331	0.9590	0.0030	0.9527	0.9645
9	9	972	0	349	0.9590	0.0030	0.9527	0.9645
9	10	623	1	298	0.9570	0.0036	0.9493	0.9636
10	10	324	0	323	0.9570	0.0036	0.9493	0.9636
10	11	1	0	1	0.9570	0.0036	0.9493	0.9636

Source: Authors.

per 1,000. Stata also produces standard errors for the cumulative survival function, which are the standard errors for the IMR and U5MR. For example, the standard error for the IMR is 2.2 per 1,000, or 8.9 percent of the IMR (known in this form as the relative standard error). Figure 3.1 is the chart produced by the `gr` option in the `ltable` command above—it shows the cumulative survival function and the 95 percent confidence intervals around the point estimates.

Disaggregated analysis and sample weights

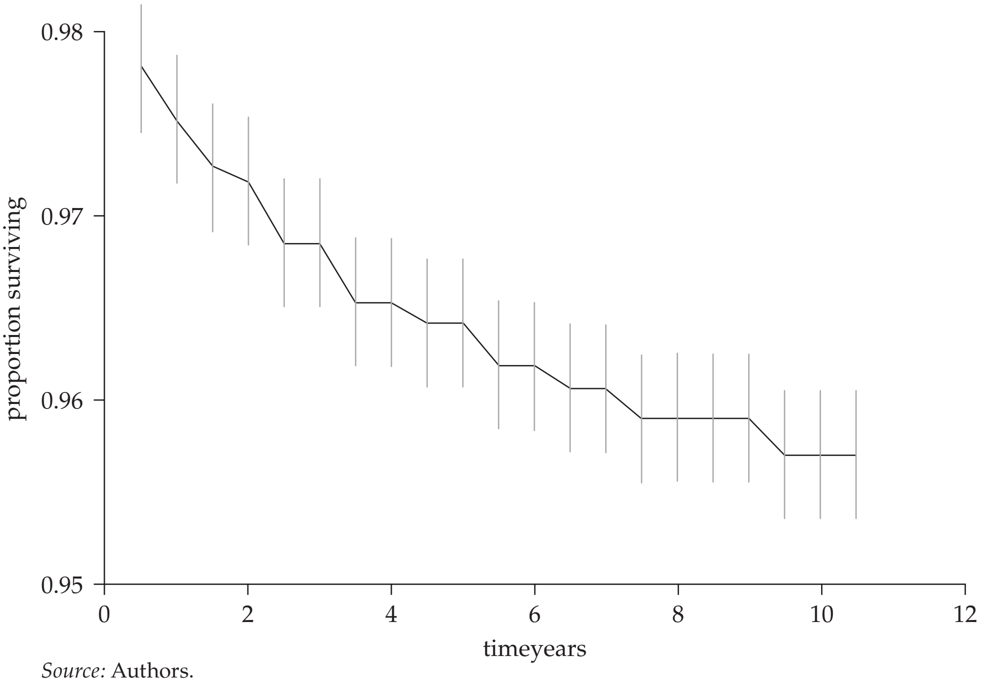
Disaggregated analysis—e.g., mortality rates for different wealth groups—is easily undertaken simply by adding to the `ltable` command a `by(`*`variable name`*`)` option. This produces as many life tables as there are categories in the stratifying variable.

Sampling weights cannot be used with `ltable`. Weighted samples can be handled using Stata's `st` commands. When the data are declared to be survival data with the `stset` command, `pweights` can be specified. The `sts list` command can be used to generate the life table from which the IMR and U5MR can be read.[2]

---

[2] For a brief discussion of the difference between `ltable` and `sts list`, see http://www.stata.com/support/faqs/stat/ltable.html.

Figure 3.1 *Survival Function with 95 Percent Confidence Intervals, Vietnam, 1988–98*



Source: Authors.

Incomplete fertility history and indirect mortality estimation

The incomplete fertility history necessitates the use of the indirect method of mortality estimation (United Nations 1983). This involves superimposing a model life table on data from the incomplete fertility history identifying (i) the number of children born to each woman and (ii) the number surviving. Computation can be done using the `DOS` program QFIVE (United Nations 1983). This section outlines the steps involved using a worked example based on the 1993 South Africa LSMS (Wagstaff 2000).

Preparing the data for QFIVE

QFIVE requires data for each of seven age groups on (i) the number of women in the sample, (ii) the number of children born, and (iii) the number of children surviving. The age groups are: 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, and 45–49. These summary data can be computed in Stata, but then need to be placed into a text file that can be read into QFIVE. Typically, the age of the women from whom the fertility history is collected is not recorded in the fertility module but needs to be merged from the household roster. The first step is to sort the household roster by the individual ID. In the commands immediately below *`filename`* is the name of the household roster file; `hhid` is the household ID, and `pcode` is the personal ID in the household.

```
use filename
sort hhid pcode
save filename , replace
```

The next step is to merge the age data from the household roster into the fertility data file. In the `use` command below *`filename`* is the name of the fertility data file; in the `merge` command *`filename`* is the name of the household roster data file. The cases from the household roster for which there are no fertility histories (men, children, and women not interviewed) are dropped using the `keep` command. Cases outside the seven age categories used in QFIVE are dropped, as are any men inexplicably left in the data file.

```
clear
use filename
sort hhid pcode
merge hhid pcode using filename
keep if _merge==3
gen agecat=age
recode agecat 0/14=0 15/19=1 20/24=2 25/29=3 30/34=4 35/39=5
40/44=6 45/49=7 50/max=8
drop if agecat<1
drop if agecat>7
drop if gender_n ~=2
```

In the South Africa LSMS, women reporting no pregnancies are assigned a value of –2 for the number of births, `no_birth`, and for the number of children alive, `no_alive`. These are recoded zero. The `tabstat` commands then summarize, for each age group, the number of women, the number of children born, and the number of children alive.

```
gen numbrths = no_birth
recode numbrths -2=0
gen numalive = no_alive
recode numalive -2=0
tabstat numbrths numalive , by(agecat) stats(co)
tabstat numbrths numalive , by(agecat) stats(su)
```

The results then need to be inserted into a text file. This can be done manually or by pasting the Stata output into Microsoft Excel, transposing the data (using Paste Special), and copying the transposed data into Microsoft Word. The file needs to be saved as text (txt) file and set out along the lines indicated below. The first row of numbers represents the number of births in each age category, the second the number of children surviving, and the third the number of women. The spacing between the numbers is crucial. It is easiest to check the tab characters and spaces boxes on the View menu under the Tools Options menu in Word, replace the tabs with spaces, and manually line up the numbers by inserting the appropriate number of spaces. Some trial and error is inevitable here, and it is essential to check that QFIVE is reading the data correctly. The "6" in the first line refers to the month in which the data were collected; "1993" is the year; "3" indicates both boys and girls; and "1" indicates that the data refer to the number of women, the number of children born, and the number of children surviving.

Input text file for `QFIVE`

```
S Africa
 6 1993 3    1
      312     1608     2841     3948     4294     4391     3734
      290     1469     2554     3490     3744     3794     3186
     2034     2063     1683     1479     1250     1099      853
```

Obtaining and interpreting output from QFIVE

Select option 1 (Enter or modify input data) in QFIVE, read in the data, and check that they are being read correctly (use PageDown to see the data). Then select option 2 (Run Q5), and ask for the data to be directed to the printer, screen, or file, as desired. If the latter option is selected, the output needs reformatting to be legible: read the output file into Microsoft Word (select Plain Text as encoding if prompted), select the entire file (Ctl-A), choose a size 8 font, select landscape as the orientation from Page Setup, and select Page Width as the zoom level. This will allow the output to be viewed without wrapping on both screen and paper.

In its output, QFIVE first reproduces the input data (see table 3.2). Then it produces estimates of the IMR, U5MR, and child mortality (mortality between the ages of 1 and 5). It produces separate estimates for each age group and for each of eight different model life tables. The four shown in table 3.3 are based on the popular Coale-Demeny life tables.[3] The estimates indicate how mortality rates vary with the age of the cohort—mostly rates are higher the older the cohort. QFIVE indicates the estimated date to which the mortality rate refers, so the estimated rates for different years (see figure 3.2) can be reported (and graphed). Or an average over cohorts can be calculated. If this is done, it is usual to ignore the rates for the two youngest and two oldest age groups on the grounds that they reflect births to young women who are unrepresentative and births that took place more than 10 years before the survey. One option is to take a simple average of the rates occurring in the age groups 25–29, 30–34, and 35–39, though sometimes it may make sense to include at least the rate for the 20–24 age group as well. If the latter group is ignored, the results in

Table 3.2 *QFIVE's Reproduction of Input Data for South Africa*

```
INPUT DATA FOR S Africa
BOTH SEXES
ENUMERATION DATE: JUN 1993

  Age Group          Number          Number of          Number of
     of                of            Children           Children
    Women            Women           Ever Born          Surviving

    15-19            2034.              312.               290.
    20-24            2063.             1608.              1469.
    25-29            1683.             2841.              2554.
    30-34            1479.             3948.              3490.
    35-39            1250.             4294.              3744.
    40-44            1099.             4391.              3794.
    45-49             853.             3734.              3186.

   MEAN AGE AT MATERNITY WAS NOT GIVEN. THE DEFAULT VALUE OF 27.0 WILL
BE USED.
```

Source: Authors.

---

[3] `QFIVE` also produces mortality rates for the UN's own life tables; the results are not shown here.

Table 3.3 *Indirect Estimates of Child Mortality, South Africa*

COALE-DEMENY:	NORTH		SOUTH		EAST		WEST	
AGE OF WOMAN	REFERENCE DATE	q	REFERENCE DATE	q	REFERENCE DATE	q	REFERENCE DATE	q
INFANT MORTALITY RATE: q(1)								
15–19	1992.2	.065	1992.2	.063	1992.2	.070	1992.2	.068
20–24	1991.0	.069	1991.0	.076	1990.9	.079	1990.9	.075
25–29	1989.3	.071	1989.2	.081	1989.1	.085	1989.2	.079
30–34	1987.4	.074	1987.2	.087	1987.0	.092	1987.1	.084
35–39	1985.2	.075	1984.9	.091	1984.7	.097	1984.9	.086
40–44	1982.8	.073	1982.3	.092	1982.1	.097	1982.4	.085
45-49	1980.0	.071	1979.2	.092	1978.9	.097	1979.5	.084
PROBABILITY OF DYING BETWEEN AGES 1 AND 5: $_4q_1$								
15–19	1992.2	.039	1992.2	.018	1992.2	.018	1992.2	.027
20–24	1991.0	.042	1991.0	.026	1990.9	.022	1990.9	.031
25–29	1989.3	.043	1989.2	.031	1989.1	.025	1989.2	.034
30–34	1987.4	.046	1987.2	.036	1987.0	.029	1987.1	.038
35–39	1985.2	.047	1984.9	.040	1984.7	.031	1984.9	.039
40–44	1982.8	.045	1982.3	.041	1982.1	.031	1982.4	.039
45–49	1980.0	.043	1979.2	.041	1978.9	.031	1979.5	.038
PROBABILITY OF DYING BY AGE 5: q(5)								
15–19	1992.2	.101	1992.2	.079	1992.2	.086	1992.2	.093
20–24	1991.0	.108	1991.0	.100	1990.9	.099	1990.9	.104
25–29	1989.3	.111	1989.2	.109	1989.1	.108	1989.2	.110
30–34	1987.4	.116	1987.2	.120	1987.0	.118	1987.1	.118
35–39	1985.2	.119	1984.9	.128	1984.7	.124	1984.9	.122
40–44	1982.8	.115	1982.3	.128	1982.1	.125	1982.4	.121
45–49	1980.0	.112	1979.2	.130	1978.9	.125	1979.5	.119

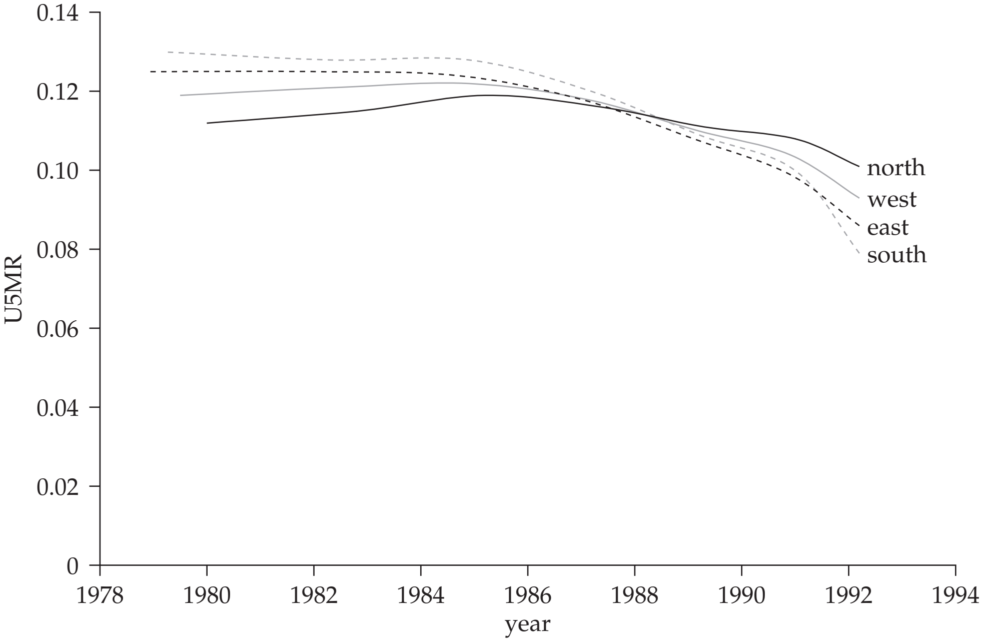
```
NOTE: A q VALUE OF .999 DENOTES VALUE BELOW A LEVEL 1 MODEL LIFE TABLE.
           "      .000       "       ABOVE A LEVEL 25        "
```

Source: Authors.

table 3.3 point toward a U5MR for South Africa of 115–119 per 1,000 for 1987. Hill and Yazbeck (1994) and Hill et al. (1999) provide guidance for a large number of developing countries on which one of the Coale-Demeny life tables is most appropriate. In this case, if the North model is used, a U5MR (for 1987) of 115 per 1,000 is obtained.

Disaggregated analysis—for example, mortality rates for different wealth groups—is easily undertaken by computing the necessary input data for QFIVE separately for each subgroup. Sample weights can be handled by specifying the sample weighting scheme when obtaining the data in Stata for the QFIVE input.

Figure 3.2 *Indirect Estimates of U5MR, South Africa*



Source: Authors.

References

Gwatkin, D. R., S. Rustein, K. Johnson, R. Pande, and A. Wagstaff. 2000. "Socioeconomic Differences in Health, Nutrition and Population." World Bank Health, Nutrition and Population discussion paper. World Bank, Washington, DC.

Hill, K., and A. Yazbeck. 1994. "Trends in Child Mortality, 1960–90: Estimates for 84 Developing Countries." World Development Report 1993, Background Paper No. 6. World Bank, Washington, DC.

Hill, K., R. Pande, M. Mahy, and G. Jones. 1999. *Trends in Child Mortality in the Developing World: 1960 to 1996.* New York: UNICEF.

International Monetary Fund, OECD, United Nations, and World Bank. 2000. *A Better World for All: Progress Towards the International Development Goals.* Washington, DC: International Monetary Fund, OECD, United Nations, and the World Bank.

UNICEF. 2001. *Progress Since the World Summit for Children: A Statistical Review.* New York: UNICEF.

United Nations. 1983. *Indirect Techniques for Demographic Estimation.* New York: United Nations.

Wagstaff, A. 2000. Socioeconomic Inequalities in Child Mortality: Comparisons across Nine Developing Countries. *Bulletin of the World Health Organization* 78(1): 19–29.

Wagstaff, A., and N. N. Nguyen. 2003. "Poverty and Survival Prospects of Vietnamese Children under Doi Moi." In *Economic Growth, Poverty and Household Welfare: Policy Lessons from Vietnam,* ed. P. Glewwe, N. Agrawal, and D. Dollar. Washington, DC: World Bank.