Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Health Statistics 2

Towards good practice for health statistics: lessons from the Millennium Development Goal health indicators

Christopher J L Murray

Lancet 2007; 369: 862–73

This is the second in a Series of four articles about health statistics

Harvard University, School of Public Health and Initiative for Global Health, Cambridge, MA, USA (Prof C L Murray MD)

Correspondence to:
Prof Christopher J L Murray
christopher_murray@harvard.edu

Health statistics are at the centre of an increasing number of worldwide health controversies. Several factors are sharpening the tension between the supply and demand for high quality health information, and the health-related Millennium Development Goals (MDGs) provide a high-profile example. With thousands of indicators recommended but few measured well, the worldwide health community needs to focus its efforts on improving measurement of a small set of priority areas. Priority indicators should be selected on the basis of public-health significance and several dimensions of measurability. Health statistics can be divided into three types: crude, corrected, and predicted. Health statistics are necessary inputs to planning and strategic decision making, programme implementation, monitoring progress towards targets, and assessment of what works and what does not. Crude statistics that are biased have no role in any of these steps; corrected statistics are preferred. For strategic decision making, when corrected statistics are unavailable, predicted statistics can play an important part. For monitoring progress towards agreed targets and assessment of what works and what does not, however, predicted statistics should not be used. Perhaps the most effective method to decrease controversy over health statistics and to encourage better primary data collection and the development of better analytical methods is a strong commitment to provision of an explicit data audit trail. This initiative would make available the primary data, all post-data collection adjustments, models including covariates used for farcasting and forecasting, and necessary documentation to the public.

Health statistics, often viewed as a dry, dull necessity, are at the centre of several worldwide health controversies. Five factors are fuelling the tension between the supply and demand for high quality health information. First, the need for greater accountability and transparency from governments and international agencies is increasing the demand. Civil society groups, the donor community, scientists, and the public want to benchmark progress and performance of public health and medicine. An important example of this tendency is the emphasis on monitoring, including the health-related Millennium Development Goals (MDGs)[1,2] and the creation of the Healthcare Commission in the UK to independently monitor the National Health Service (NHS).[3]

Second, the media, civil society, and the general public are more sceptical about both statistical and scientific claims.[4] An example is the public dismay about the confusing messages on fat in the diet over the past two decades.[5] Evidence of government manipulation of data during the severe acute respiratory syndrome (SARS) epidemic added to the decline in trust.[6] Third, many representatives of the technical and scientific community and the general public are becoming increasingly sophisticated consumers of information. The scope of relevant information is expanding from simple descriptive epidemiology about health to dimensions of public health and medicine such as quality, efficiency, and equity. Consumers often need more detail including quantification of uncertainty. Increased communication and access to different views through the internet are driving scepticism and sophistication, which in turn adds to the broad demand for transparency and accountability.

Fourth, as outlined in the first paper in this series,[7] leaders of various global-health programmes including WHO, many public-private global health initiatives such as the Global Fund to fight AIDS, Tuberculosis, and Malaria, and other development agencies, feel the imperative to produce more information on their programmes and the outcomes of their investments for public consumption. Yearly reports, websites, and other publications are regarded as necessary for sustaining political and financial support for their programmes, and for maintaining political priority with governments in developing countries. One result of this response has been an explosion of proposed indicators that should be measured. For example, the Drug Action Programme at WHO has 98 indicators for monitoring structure, process, and outcome[8] and the HIV department has 35 indicators.[9,10] An internal review presented to the WHO Director General in 2002 found that WHO recommended 3500 indicators covering all programme areas of the organisation. For most of these indicators, no measurement strategy has been proposed and no measurements have been produced.

Fifth, as demand for health information grows, primary data collection platforms in most developing countries are not rapidly improving.[7] However, the information technology revolution has not yet had a major effect on platforms for primary data collection in health systems in most developing countries.[11]

In this paper, I explore good practice for health statistics from a worldwide and national perspective. Although there are many dimensions to improvement of health statistics, I concentrate on three issues that are

central to fostering good practice in the use of health statistics: focusing on priority indicators; correct use of crude, corrected, or predicted statistics; and the need for explicit data audit trails. Taken together, addressing these issues would catalyse better worldwide health statistics practice and stimulate increased national interest in strengthening fundamental data platforms. I believe that increased focus on the production, analysis, and dissemination of priority health indicators will create demand for valid, reliable, and comparable health information, which will stimulate national efforts to strengthen platforms for primary data collection. I do not discuss in detail the interventions that might be necessary to strengthen such platforms. The MDG health-related indicators are used throughout the paper to draw attention to the difficulties with present indicators in terms of conceptualisation, implementation, and measurement. Although the MDG health-related indicators have been developed for high-level policy use, they represent the issues that generally apply to priority health indicators. Technical terms used in this paper are explained in the panel.

Focusing on priority indicators

With thousands of indicators recommended but few measured well, the worldwide health community needs to focus its efforts on improving measurement of a small set of priority areas. Prioritisation of indicators is important for two reasons. The first is cost. Human resources in the measurement field are extremely scarce, both nationally and internationally. The second is visibility. Indicators drive policy attention and resources nationally and locally. This inevitable dynamic means that health problems with priority indicators will receive more attention than those that are not measured or not measured as well. Prioritisation requires several questions to be answered.

What is the proposed indicator intended to measure?

Indicators can be classified into six categories on the basis of what it is they measure: health outcomes, risk factors, intervention coverage, structure, process, and non-health-related results. All these types of indicators have important uses in different contexts. Part of the challenge in assessment of indicators is to understand at what level they will be used. At the highest level, such as the MDG indicators, the audience is extremely broad including the technical community, governments, and the general public.

What is the public-health significance of the indicator?

Public-health importance is probably greatest for indicators of health outcomes, intervention coverage, and perhaps risk factors. Many health outcomes are of public-health importance, but because health problems are changing with new challenges, such as the epidemic of non-communicable diseases in developing countries and emergence of pandemic influenza, assessment of public-health significance must be regularly revisited. Coverage of interventions is also important because only through the delivery of effective interventions to those in need can health outcomes be improved. Finally, risk factors, such as tobacco consumption, can be so strongly linked to health outcomes that they are effectively measures of future health outcomes in

***Panel:* Definitions of technical terms**

Indicator
A variable measured to monitor progress or assess what works and what does not.

Validity
Validity refers to the extent to which a measurement is capturing what it is intended to measure. There are different types of validity such as face validity, content validity, criterion validity (denoting predictive validity and concurrent validity), and construct validity (denoting convergent and discriminant validity).

Reliability
Reliability refers to the repeatability or consistency of a set of measurements or measuring instrument; for example, test-retest reliability where a test and a retest are compared.

Comparability
Measurements are comparable if the same value means the same thing in the settings being compared. Two thermometers, one in Farenheit and one in Celsius, can both be valid and reliable but they do not give comparable results.

Out-of-sample
Prediction about ranges of values that are not in the investigator's sample (ie, that the investigator's data set does not cover).

Out-of-time
Prediction about individuals, populations, etc, in time outside the time range of the investigator's sample.

Forecasting
Forecasting is the process of estimation in unknown situations. Predicting is a more general term and connotes estimating for any time series, cross-sectional, or longitudinal data. Forecasting is commonly used when discussing time series data.

Farcasting
Farcasting is trying to predict the value of a variable in a place that may be far away but is not a future value.

Prior
The prior is a reflection of some information the investigator has before the observations in the data set (the investigator should state explicitly how the information on the prior was obtained). The prior is the sum of what is known about the relationship under study.

	Type of indicator	Public-health importance of the indicator	Measurement strategy	Data availability				Predominant type of statistic
				1990	2000	2003	1990–2005	
4. Prevalence of underweight children under 5 years of age								
4a. Children under 5 moderately or severely underweight, percentage*	Risk factor	High	Household surveys with anthropometric measurements with some inconsistency of age groups measured	8%	31%	5%	8%	Corrected
4b. Children under 5 severely underweight, percentage	Risk factor	Adds little value to 4a	Household surveys with anthropometric measurements, some inconsistency of age groups measured	0%	29%	3%	5%	Corrected
5. Proportion of population below minimum level of dietary energy consumption								
5a. Undernourished as percentage of total population*	Risk factor	Low; adds little to prevalence of underweight	Details are not available	0%	0%	0%	13%	Predicted
5b. Undernourished, number of people	Risk factor	Low; same information content as 5a	Details are not available	0%	0%	0%	13%	Predicted
13. Under-5 mortality rate*	Health outcome	High	Vital registration in countries with complete systems, complete birth histories, or children ever born and children surviving questions on household surveys	98%	100%	98%	25%	Corrected and predicted
14. Infant mortality rate*	Health outcome	Low; redundant, correlation coefficient with under-5 mortality rate in 2000 is 0·99	Vital registration in countries with complete systems, complete birth histories, or children ever born and children surviving questions on household surveys	99%	100%	100%	26%	Corrected and predicted
15. Proportion of 1 year-old children immunised against measles*	Intervention coverage	Medium/low; represents less than 10% of the intervention package for child survival	Health service provider registries in the public sector, household surveys	84%	98%	98%	83%	Crude and corrected
16. Maternal mortality ratio*	Health outcome	High	Vital registration in countries with complete systems, sibling histories collected in household surveys	82%	100%	0%	17%	Predicted
17. Proportion of births attended by skilled health personnel*	Intervention coverage	High; but not the only intervention needed to reduce maternal mortality	Household surveys, definition of skilled varies across countries	0%	34%	7%	6%	Corrected
18. HIV prevalence among pregnant women aged 15–24 years								
18a. AIDS estimated deaths	Health outcome	High	Vital registration in countries with complete systems, modeling of mortality based on estimated seroprevalence in other countries	0%	0%	67%	8%	Predicted
18b. HIV prevalence rate, aged 15–49, percentage	Health outcome	High	Antenatal clinic (ANC) serosurveillance in sentinel sites, household serosurveys. ANC sero-surveillance appears to overestimate population prevalence	0%	0%	78%	9%	Corrected and predicted
18c. HIV/AIDS prevalence rate for pregnant women 15–24 attending antenatal care in clinics in capital city*	Health outcome	Low; represents only a partial fraction of national prevalence, 18b is the true quantity of interest	Capital city ANC serosurveillance, because of variability in representativeness of sentinel clinics and demographic significance of capital city, comparability limited	0%	5%	5%	29%	Crude
18d. HIV/AIDS prevalence rates, men, estimated from national population surveys	Health outcome	Input to accurate measurement of 18b	Household serosurveys	0%	0%	3%	0%	Corrected
18e. HIV/AIDS prevalence rates, women, estimated from national population surveys	Health outcome	Input to accurate measurement of 18b	Household serosurveys	0%	0%	3%	0%	Corrected
19. Condom use to overall contraceptive use among currently married women aged 15–49 years*	Intervention coverage	Low; not a good measure of condom use in high-risk sexual intercourse	Household surveys	6%	29%	4%	8%	Corrected
19a. Condom use, men, aged 15–24 years at last high-risk sex*	Intervention coverage	Medium; condom use for any age-group for high-risk sex would be the quantity of interest	Household surveys but validity of reported rates of high-risk sex not established	0%	6%	3%	1%	Corrected
19b. Condom use, women, aged 15–24 years at last high-risk sex*	Intervention coverage	Correlation coefficient with 19a in 2000 is 0·85	Household surveys but validity of reported rates of high-risk sex not established	0%	7%	3%	1%	Corrected

(Continues on next page)

(Continued from previous page)

19c. HIV knowledge, men aged 15–24 years who know that a healthy-looking person can transmit HIV	Intervention coverage	Low; small component of 19e	Household surveys	0%	7%	0%	27%	Corrected
19d. HIV knowledge, men aged 15–24 years who know that a person can protect himself from HIV infection by consistent condom use	Intervention coverage	Low; small component of 19e	Household surveys	0%	4%	0%	1%	Corrected
19e. HIV knowledge, men aged 15–24 years with comprehensive correct knowledge of HIV/AIDS, percentage*	Intervention coverage	Low; poorly established link and partial relationship to unsafe sexual practices	Household surveys	0%	5%	3%	1%	Corrected
19f. HIV knowledge, women aged 15–24 years who know that a healthy-looking person can transmit HIV	Intervention coverage	Low; small component of 19e	Household surveys	0%	32%	0%	3%	Corrected
19g. HIV knowledge, women aged 15–24 years who know that a person can protect himself from HIV infection by consistent condom use	Intervention coverage	Low; small component of 19e	Household surveys	0%	37%	0%	3%	Corrected
19h. HIV knowledge, women aged 15–24 years with comprehensive correct knowledge of HIV/AIDS, percentage*	Intervention coverage	Low; poorly established link and partial relationship to unsafe sexual practices	Household surveys	0%	27%	3%	2%	Corrected
19i. Contraceptive use among currently married women aged 15–49 years, any method, percentage*	Intervention coverage	Low; unclear relationship to preventing HIV transmission	Household surveys	6%	29%	5%	8%	Corrected
19j. Contraceptive use among currently married women aged 15–49 years, condom, percentage	Intervention coverage	Low; less relationship to transmission potential than 19a or 19b	Household surveys	7%	29%	4%	8%	Corrected
19k. Contraceptive use among currently married women aged 15–49 years, modern methods, percentage	Intervention coverage	Low; weak relationship to decreasing HIV transmission	Household surveys	5%	29%	4%	8%	Corrected
20. Ratio of school attendance of orphans to school attendance of non-orphans aged 10–14 years								
20a. AIDS orphans (one or both parents), currently living	Non-health outcome	Low for public health but could be important for HIV related social policy	Modelled relationships based on estimated HIV seroprevalence and mortality	0%	0%	25%	3%	Predicted
20b. Orphans (both parents) aged 10–14 school attendance rate as % of non-orphans attendance rate, where HIV is >1%*	Non-health outcome	Low; not HIV specific, in nearly all countries result is 100%	Household surveys	0%	27%	3%	3%	Corrected
21. Prevalence and death rates associated with malaria								
21a. Malaria death rate per 100 000, ages 0–4 years*	Health outcome	High	Vital registration in countries with complete systems; in nearly all endemic countries, based on verbal autopsy data for demographic surveillance sites, or epidemiological models	0%	100%	0%	6%	Predicted
21b. Malaria death rate per 100 000, all ages*	Health outcome	Low; death rates over the age 0–4 are very low	Vital registration in countries with complete systems; in nearly all endemic countries, based on verbal autopsy data for demographic surveillance sites, or epidemiological models	0%	100%	0%	6%	Predicted
21c. Malaria prevalence, notified cases per 100 000 population*	Health outcome	Low; notified cases are not a measure of prevalence	Administrative data collected at public facilities	0%	55%	0%	4%	Crude
22. Proportion of population in malaria-risk areas using effective malaria prevention and treatment measures								
22a. Malaria prevention, use of insecticide-treated bed nets in population <5, percentage*	Intervention coverage	High	Household surveys	0%	18%	3%	2%	Corrected

(Continues on next page)

(Continued from previous page)								
22b. Malaria treatment, percentage of population <5 with fever being treated with antimalarial drugs*	Intervention coverage	Moderate; resistance makes 'effective antimalarial drugs' better indicator	Household surveys, validity not established.	0%	18%	4%	2%	Corrected
23. Prevalence and death rates associated with tuberculosis								
23a. Tuberculosis death rate per 100 000*	Health outcome	High	Vital registration in countries with complete vital registration, models for all other countries	98%	98%	98%	31%	Predicted
23b. Tuberculosis prevalence rate per 100 000 population*	Health outcome	High	No measurement strategy; modelled estimates based on case-notifications	98%	98%	98%	31%	Predicted
24. Proportion of tuberculosis cases detected and cured under directly observed treatment success (DOTS)								
24a. Tuberculosis, DOTS detection rate, percentage*	Intervention coverage	High	Health service provider registries for detected cases, no measurement strategy for denominator	0%	63%	95%	40%	Predicted
24b. Tuberculosis, DOTS treatment success, percentage*	Intervention coverage	High	Health service registries	0%	76%	92%	42%	Corrected
25. Proportion of population using solid fuels*	Risk factor	Moderate; real quantity of interest is indoor air pollution	Household surveys	1%	6%	61%	6%	Predicted
30. Proportion of population with sustainable access to an improved water source, urban and rural								
30a. Water, percentage of population with access to improved drinking water sources, rural*	Risk factor	Moderate; not clear that separate urban and rural indicators necessary	Household surveys: some issues in the consistent definition of "improved"	7%	0%	0%	11%	Predicted
30b. Water, percentage of population with access to improved drinking water sources, total	Risk factor	High	Household surveys: some issues in the consistent definition of "improved"	71%	0%	0%	11%	Predicted
30c. Water, percentage of population with access to improved drinking water sources, urban*	Risk factor	Moderate; correlation with rural in 2002 is 0·69	Household surveys: some issues in the consistent definition of "improved"	84%	0%	0%	12%	Predicted
31. Proportion of population with access to improved sanitation, urban and rural								
31a. Sanitation, percentage of population with access to improved sanitation, rural*	Risk factor	Moderate; not clear that separate urban and rural indicators necessary	Household surveys	72%	0%	0%	11%	Predicted
31b. Sanitation, percentage of population with access to improved sanitation, total	Risk factor	High	Household surveys	67%	0%	0%	10%	Predicted
31c. Sanitation, percentage of population with access to improved sanitation, urban*	Risk factor	Moderate; not clear that separate urban and rural indicators necessary	Household surveys	76%	0%	0%	11%	Predicted
46. Proportion of population with access to affordable essential drugs on a sustainable basis*	Intervention coverage	High	No measurement strategy	0%	0%	0%	0%	No data
All official MDG indicators combined				30%	44%	29%	15%	

*Official MDG indicators. All others are official supplemental series.

Table 1: **MDG health-related indicators (number and name) reported to the General Assembly and officially reported supplemental series**

themselves. Not all outcomes or all intervention-coverage indicators will meet a public-health importance criterion.

Table 1 provides my assessment of the public-health importance of every health-related MDG indicator including official supplemental indicators. The analysis in table 1, including data availability for each indicator, is strictly based on the official UN MDG website;[12] there could well be other surveys and analyses that are not used for MDG monitoring. The 45 indicators in the table could probably be decreased to 16 that have high public-health importance and contribute unique information. For example, for a set of priority indicators, overall coverage of clean water is probably enough instead of urban and rural breakdowns. Arguably, urban and rural breakdowns do not provide a deep insight into differences in clean water inequalities across countries. A more useful indicator for this comparison

might be the amount of clean water coverage in the bottom wealth quintile.

How well does the indicator measure the quantity of interest?

Since indicators can drive policy attention, resources, and implementation, every effort should be made to measure or at least estimate the real quantity of interest. For health-outcome indicators, does the proposed indicator measure the actual health outcome, a component of the outcome, or a proxy that is believed to be correlated to health outcome? In table 1, indicator 18c "HIV/AIDS prevalence rate for pregnant women 15–24 attending antenatal care in clinics in capital city", could be an important measure for the local HIV programme in the capital city, but this indicator is a poor measure of the real quantity of interest in the MDG framework—ie, national prevalence. Use of partial or proxy measures runs the danger that they slowly become the de facto quantity of interest in policy debates, distracting attention from the original objective of a policy or programme. Even when an indirect or proxy measure is the only option, mapping from the units of that measure into units of the quantity of interest is preferable—eg, ratios of children surviving to children ever born mapped into child mortality.

For intervention coverage indicators, the ideal is to measure the proportion of potential health gain that is delivered for a specific health problem.[13] The target for MDG4 is to reduce child mortality; the ideal intervention coverage would measure the proportion of child mortality reduction that could be achieved through the whole collection of affordable interventions that are delivered. Compared with this standard, measurement of only one intervention, measles immunisation in the MDGs, is capturing less than 5% of what can be delivered. If only one intervention is to be measured, preference should be given to the intervention that would have the biggest effect on the outcome if delivered. Because intervention coverage indicators have the potential to drive managerial attention and resource allocation, claims that part indicators such as measles immunisation are good proxies for the rest of child survival interventions are unproven. The potential for indicator-driven policy also argues for the use of composites capturing the coverage of a set of interventions targeting a health problem.

Is the indicator value readily interpretable?

There should be a monotonic relation between the value of the indicator and what is desired (all other things being equal). The child death rate is a monotonic indicator; lower rates are always better than are higher ones. However, the caesarean section rate is not a monotonic indicator. Too low and too high rates are equally bad, and there is no consensus on the optimum rate.

Is there a practical measurement strategy?

An indicator should not be a priority unless a measurement strategy that will produce valid, reliable, timely, and comparable measurements has been developed. This tenet does not mean that a measurement strategy has been implemented, otherwise, we would be trapped measuring only what has already been measured. Rather, it means that a plan has been developed. Although validity and reliability are familiar notions, comparability and timeliness have often not received adequate attention in the development of measurement strategies. Comparability is crucial—why measure an indicator over time and across countries if the measurements cannot be compared? Murray and colleagues[14] emphasise the difference between validity and comparability by noting that two thermometers, one in Celsius and one in Fahrenheit, can both be valid and reliable but 26 degrees on each is not comparable.

Unfortunately only nine of the 16 high-importance MDG indicators have a reasonably well developed measurement strategy that will yield valid, reliable, and comparable measurements. Table 1 also emphasises that data availability overall for these health-related indicators is extremely poor, ranging from 0% for access to essential drugs to 100% for tuberculosis rate, although tuberculosis rate is almost entirely predicted from a model. For the health-related MDG indicators, overall availability of any type of statistics is 15% for 1990–2005. The 15% figure refers only to developing countries for all of the health-related MDG indicators. Full analysis of existing data sources in countries would probably expand this availability, but existing data are often not fully used within countries or for worldwide comparative studies.

How should equity dimensions of an indicator be captured?

Gwatkin[15] has argued that average levels for an indicator can mask widening inequalities. To measure inequalities over time, disaggregation is necessary. Ideally, the total inequality of a health indicator across the population could be measured.[16] Total inequality can be thought of as between-group inequality plus within-group inequality.[17] For many indicators, however, only between-group inequality can be readily measured. The choice of groups for indicator disaggregation, such as rich and poor, urban and rural, occupation, race, or ethnic origin could profoundly change the comparisons of inequality across populations or over time. The disaggregation of an indicator that is proposed for monitoring inequalities should accord with evidence that this disaggregation captures the largest fraction of variance in the indicator across the populations being analysed.

Correct use of crude, corrected, or predicted statistics

Boerma and colleagues[7] have discussed the platform for primary data collection that provides the basis for all health statistics: essential registration systems, sample

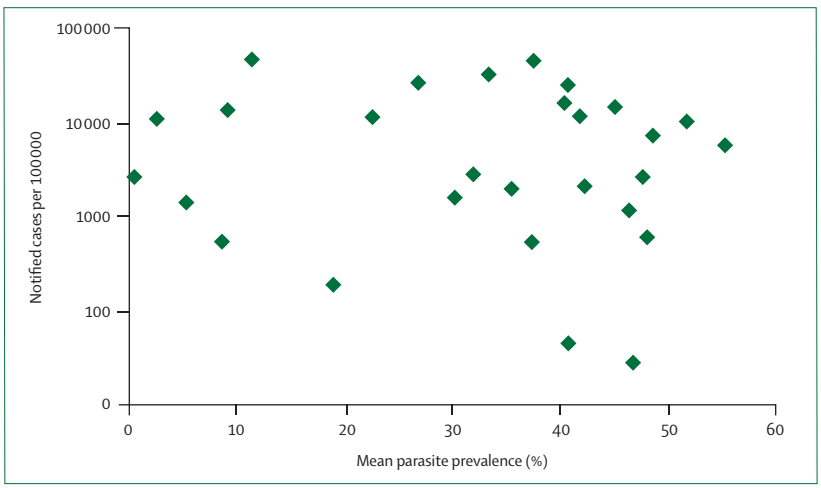

Figure 1: **MDG indicator malaria prevalence versus average parasite seroprevalence from MARA systematic review for selected sub-Saharan African countries**

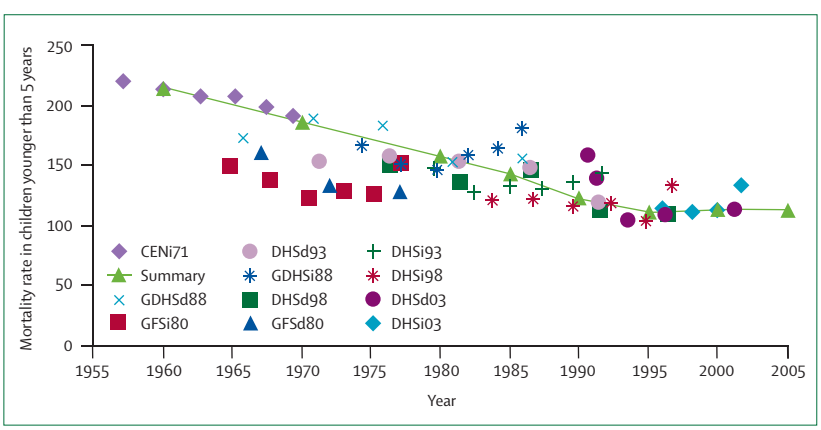

Figure 2: **All available empirical estimates of mortality rates for children younger than 5 years for Ghana and best estimates**

CENi=1971 Census, indirect. DHSd93=1993 Demographic and Health Survey, direct. DHSi93=1993 Demographic and Health Survey, indirect. DHSd98=1998 Demographic and Health Survey, direct. DHSi98=1998 Demographic and Health Survey, indirect. DHSd03=2003 Demographic and Health Survey, direct. DHSi03=2003 Demographic and Health Survey, indirect. GDHSd88=1988 Ghana Demographic and Health Survey, direct. GDHSi88=1988 Ghana Demographic and Health Survey, indirect. GFSd80=1980 Ghana Fertility Survey, direct. GFSi80=1980 Ghana Fertility Survey, indirect.

or sentinel registration systems, household surveys, censuses, budgets, and data collected by health-service providers. Primary data collection is only the first step in the generation of valid, reliable, and comparable statistics. The subsequent steps such as correcting for known biases or predicting out of sample or out of time can generate many different measurements. Health statistics can be divided into three types: crude, corrected, and predicted.

Crude health statistics

Crude health statistics are the measurements of indicators that come directly from primary data collection with no adjustments or corrections. These figures are subject to many drawbacks, including incomplete ascertainment, non-representativeness, instrument bias, misclassification, and distortion. Incomplete ascertainment or incomplete coverage is a crucial difficulty for vital registration systems and data from health-service providers.[18] Poor people and other disadvantaged groups often have the greatest health problems and do not get captured in these systems. For household surveys or other sampled data, non-representativeness can be a profound difficulty. For example, monitoring of HIV seroprevalence in antenatal clinics is often undertaken in clinics with known high prevalence.[19] For many measurements and self-reported items on surveys, the instrument itself can be biased—eg, self-reported weights are systematically under-reported by women.[20] Where events need to be classified into categories such as deaths according to the international classification of diseases and injuries, misclassification or inconsistent classification is a common drawback.[21–23] Finally, where the stakes are high, producers of data might intentionally, or be encouraged to, distort data.

Despite their restrictions, crude health statistics are proposed and used for monitoring. Figure 1 shows the MDG indicator malaria prevalence (notified cases per 100 000) plotted against the prevalence of malaria parasites in blood as systematically reviewed by the MARA (Mapping Malaria Risk in Africa) project for sub-Saharan Africa.[24] Although the prevalence of clinical cases of malaria would not necessarily be perfectly correlated with parasitaemia, the absence of any relation even on log scale suggests that because of variations in ascertainment this MDG indicator cannot be used for any interpretation. In some cases, such as complete vital registration systems, crude statistics can be unbiased and appropriate for analytical use. Since bias in crude statistics is expected, the burden of proof should be to show that crude data are in fact unbiased before they are used for comparative purposes.

Corrected health statistics

Corrected health statistics are measurements of indicators for which two types of analytical effort might have been undertaken: mapping to the quantity of interest and correction for a range of known biases. Mapping into the quantity of interest includes all the measurements where the primary data collected are on an indirect result of the event under study. For example, census or survey data on the responses of mothers regarding the number of children they have ever borne and the number that are alive are used in many countries to calculate mortality in children younger than 5 years. This mapping is based on certain assumptions and models;[25,26] mapping into the quantity of interest introduces uncertainty because of parameter uncertainty, residual unexplained variance, and model choice.

Correction for known bias ranges from routine procedures such as use of sample weights in household survey analysis to more complex procedures that could include analytical models. Demographers use methods[27] to estimate the incompleteness of vital registration data and then apply these to obtain corrected mortality rates.

Less formal or structured methods are routinely used to correct HIV seroprevalence data in antenatal clinics to generate national estimates to address non-representativeness.[28] Correction for known bias is very important if valid, reliable, and comparable health statistics are to be generated; however, substantial scope for legitimate disagreement between analysts can be introduced. Often the details of efforts to correct for known bias are not in the public domain, which has stifled open debate on the best approaches to do such corrections. Corrections for known bias could in fact introduce more error into health statistics; therefore, open debate should be encouraged.

The most technical approach to correcting for known bias is to systematically review and use all available primary data and attempt to reconcile differences between data sources.[29–31] Systematic review and data reconciliation should not be confused with more qualitative triangulation.[32] Figure 2 provides another example of this approach by summarising all the available data on child mortality in Ghana and the corrected figures based on these data.[27,33] An additional example of this approach is the work to develop internally consistent figures for incidence, prevalence, and death in the global burden of disease project.[34,35] Software such as DISMOD II is used to identify inconsistencies between different data sources, which the analyst must then reconcile. The advantage of systematic reviews and data reconciliation is that all relevant information is used to correct for known bias. The disadvantage is that they can require substantially more analytical work.

Predicted health statistics

Predicted statistics are based on a model relating the quantity of interest to covariates. Two types of predicted statistics are widely used. The first is forecasting, whereby a relation is established during a period of observation and then used to predict out of time into the future. A common use of forecasting is to update corrected statistics to a more recent period to produce series of comparable statistics for a base year, such as the MDG indicator data on maternal mortality ratio in 2000, or many statistics in the MDG database for indoor air pollution, water supply, or tuberculosis prevalence. Predictions are also frequently used to generate figures in settings where no primary data and thus no corrected statistics are available. Since the methods are identical to those for forecasting, prediction out of sample but in the same time period has been termed farcasting. An example of farcasting is MDG estimates of solid fuel use that have been based for many countries on models relating solid fuel use to GDP per capita in those countries with data. Predicted statistics have uncertainty resulting from model choice and parameter and unexplained variance in the model.

Figure 3 provides trends for maternal mortality from 1990 to 2000 for Afghanistan, Angola, Pakistan, Nigeria, South Africa, Burundi, and Algeria. These trends are not based on any empirical measurements and are strictly from predictive models. Nevertheless, the trends are different and show striking fluctuations. In Pakistan, the maternal mortality ratio is predicted to be increasing, although in Nigeria the ratio is predicted to be declining. Since these trends are not based on any real findings, to infer different amounts of progress on this MDG target would be completely without basis. The MDG database does not provide uncertainty for the trends in this indicator or other predicted indicators.

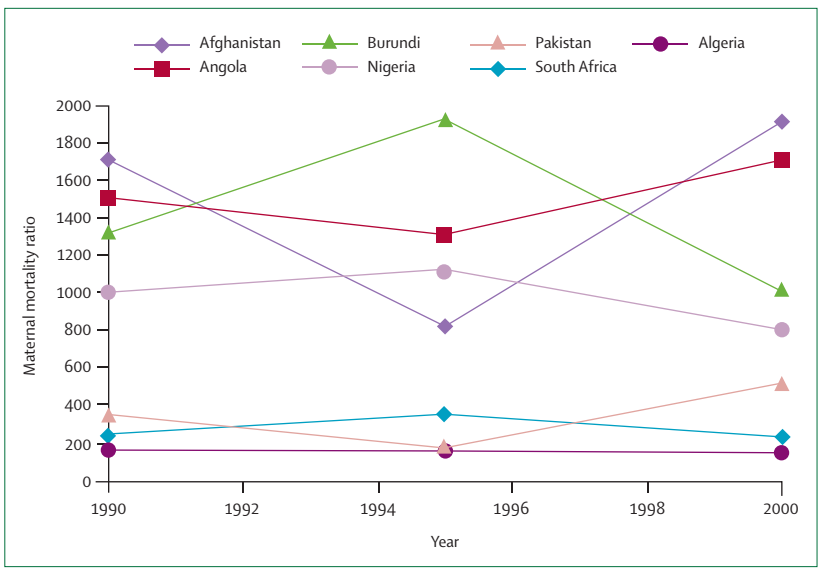

Figure 3: **Trends in maternal mortality based strictly on predicted statistics for seven countries**

There is a grey zone where the corrections to primary data are substantial enough that they can become difficult to distinguish from predicted statistics. In a Bayesian context, predictions provide priors and the primary data with correction allow these priors to be updated. There are, however, cases where the recorded data give little information beyond the prior.

When to use what

Health statistics are necessary inputs to planning and strategic decision making, programme implementation, monitoring progress towards targets, and assessment of what works and what does not. Crude statistics that are biased have no role in any of these steps. They are, however, an important input to research and are the basis for corrected statistics with greater validity, reliability, and comparability. There is no cogent reason why crude statistics with known biases should be used when corrected statistics are available or can be developed. Nevertheless, several agencies report crude statistics. For example, the US Centers for Disease Control and Prevention report on levels of obesity by state with self-reported weights and heights in the behavioural risk factor surveillance system survey, which are known to

underestimate obesity by roughly 50%.[20] By reporting crude statistics, the onus for correction of known bias to allow for interpretation is shifted to the user. Although this approach makes the task easier for the data generator, it is predicated on a strange premise: the users, including policymakers and the concerned public, are in a better position to adjust for known bias than is the health statistician. In fact, the failure to correct for known bias underlies some of the most egregious examples of health statistics.

Publication of government data in WHO's communicable disease global atlas[36] on notified cases and deaths for rabies, cholera, sexually transmitted infections, and malaria are notable examples of the promotion of crude statistics. WHO claims that the atlas "is bringing together for analysis and comparison standardized data and statistics for infectious diseases at country, regional, and global levels".[36] The style of presentation with maps encourages cross-country comparisons, but simple inspection shows that because of ascertainment bias, misclassification, and non-representativeness, the numbers are not valid, reliable, or comparable—eg, those for rabies for Iran and Pakistan differ by two orders of magnitude.

Political arguments, such as the importance of respecting national sovereignty, are often given to explain the use of crude figures. But political pedigree is not a legitimate argument to favour one statistic over another. Frequently governments do have the best primary data and corrected statistics, but the preference for data from the government must be based on the merits of the data, not the politics. The importance of distinguishing political pedigree from substantive arguments about the reliability, validity, and comparability of health statistics is just as true nationally, when geographic subunits are being compared, as it is worldwide.

For forward-looking decision making, advocacy for funds, planning, strategic decisions, research, and development investments, people who make decisions need the best available information. We cannot argue that we should allocate no resources to an area of need because good primary data and corrected statistics are not available. For these decisions, when corrected statistics are not available, predicted statistics can play an important part. The publication of the world population prospects every 2 years by the UN Population Division[37] is a good example of the use of predicted statistics. This publication provides population, mortality, and fertility statistics for all countries from 1950 to 2050. For some countries, much of the trend from 1950 up to now is based on corrected statistics, but for many countries much or all of the sequence is based on predictions. Because the world population prospects is revised every 2 years, the entire set of figures from 1950 to 2050 is revised when new data become available. Even though much of the sequence is based on predictions, these basic demographic figures are crucial for various planning and strategic decisions.

For monitoring progress towards agreed targets and assessment of what is effective and what is not, the case for using predicted statistics is unconvincing. Nationally or locally, the reason for monitoring is to find out whether present policies and programmes are leading us in the right direction or to identify when unexpected factors are changing trends. If data for monitoring are predicted and the prediction model explains much of the observed variance, the predicted statistics can correctly identify the true trend in many cases. However, the prediction is not sensitive to the actual policies and programmes that have been pursued. That statistics used for monitoring have zero probability of detecting an unexpected trend does not seem right. Imagine enrolling your child in a school, which then hands you the results of your child's examination scores for the end of the year on the first day of attendance. These results have been forecast by a very good model and might be right most of the time. However, the grades or assessment are not affected in any way by what your child does during the year.

A test for the use of statistics for monitoring and assessment should be that the statistics in question have a reasonable probability of detecting real changes in the quantity of interest. This concern is great when we intend to assess what works and what does not; here predicted statistics should have no role. Because predicted statistics have a zero probability of detecting changes from an abrupt adjustment in policy, they should not be used for this type of monitoring and assessment.

Predicted statistics, however, have been substantially used in monitoring. For example, 20 of the health-related MDG indicators in table 1 use some predicted statistics. The World Health Report in 2000 for health system performance used predicted statistics extensively.[38] Two types of argument have been made to support the use of predicted statistics for monitoring. First, several countries in greatest need will have little or no data to report, and if no statistics are reported the topic will lose policy attention. Second, worldwide advocacy needs information on global monitoring. Although the use of predicted statistics to bring attention to a difficulty seems warranted both nationally and internationally, the temptation to use such information once produced for actual monitoring and even assessment is very great. I believe that if predicted statistics are to be used in such cases, the user must make it very clear that predicted statistics are being used and this use should not become a routine activity. Therefore, publication of databases of annual predicted figures, such as for some MDG indicators, does not seem justified. Reported indicator series should, at an absolute minimum, indicate whether figures are based on crude, corrected, or predicted values. Regional or worldwide aggregates are often based on mixtures of crude, corrected, and predicted data, and the relative proportion of the component national figure based on these three types should be shown to the user.

The need for explicit data audit trails

Perhaps the most effective method by which to decrease controversy over health statistics and to encourage better primary data collection and development of better analytical methods is a strong commitment to provision of an explicit data audit trail. This method would make primary data, all post-data collection adjustments, models including covariates used for farcasting and forecasting, and necessary documentation available to the public. An explicit data audit trail would allow health statistics to be subject to the scientific principle of replication. A sceptical user should be able to reproduce every figure including all steps along the way. Decision makers, the media, or the public are unlikely to use the explicit data audit trail, but the requirement to publish this trail will over time lead to improved practice. The expectation that peers in the technical community will be able to critique methods, suggest other data sources, and even generate alternative figures will provide a powerful incentive for improved measurement.

Few in the international community are opposed to the idea of an explicit data audit trail. There are three difficulties, however, that have restricted the adoption of this approach. First, documentation of all adjustments for known bias and the use of farcasting and forecasting models is time-consuming and expensive. Second, transparency about how some estimates are generated can increase debate about the validity of figures in the short term, which is a concern for some programme managers. Third, the component of the explicit data audit trail that calls for primary data to be accessible to the public or at least to interested analysts might conflict with national privacy laws governing data collection, which is a very important issue. In some countries such as the USA, data are becoming harder to obtain. For example, the National Center for Health Statistics (NCHS) has stopped releasing mortality data for counties because of privacy concerns.

Increasing concerns about privacy are in direct conflict with another international trend towards more countries adopting freedom of information laws. In 2002, Mexico passed a sweeping freedom of information act and set up a Federal Institute for Access to Public Information about Mexico, which guarantees unprecedented access to any information held by the government.[39] In India, the Right to Information Act of 2005, seems to provide similar access to information.[40] Various other countries are pursuing similar legislation.[41,42] Balancing privacy concerns with the need for information to be in the public domain to create a culture of transparency will be a major challenge in the coming decades. Once individual identifiers such as names and addresses have been removed from survey, census, or vital registration data, privacy concerns revolve around the possibility that, with some investigation, individuals could be identified.[43] For example, perhaps only one 84-year-old woman died in Omaha county, USA, so that the analyst could with further investigation find the name of this individual and what she died of. Such a hypothetical possibility must be balanced against the damage to the community by suppression of health information that could improve population health. In the case of death data in the USA, the NCHS position is hard to understand since death certificates for all individuals including names are public records that can be obtained in every state.

Clearly, there are important legitimate concerns about data privacy. However, in some cases arguments about data confidentiality are used by institutions to avoid the release of data to other groups. Datasets are often viewed by researchers or organisations as private possessions. Data, of course, are a classic example of a general public good. There is a potential cost that could ultimately be counted in human health of keeping data from the public domain. The principle of an explicit data audit trail and replicability might also be preserved if mechanisms exist to allow restricted access to various datasets with stronger safeguards. Journals such as *The Lancet*, funding agencies, and governments all have an important part to play in transformation of the culture of data from a feudal to an open democratic model. Any efforts that move towards an explicit data audit trail should be encouraged, including intermediate steps such as provision of detailed information on primary data sources and types of adjustments that have been undertaken.

Conclusion

Several good practice recommendations for health indicators and their measurement follow from this discussion. These recommendations could lead over time to a more focused effort on production of valid, reliable, and comparable information to serve many information needs.

National and worldwide efforts to improve health statistics should focus on a smaller set of priority indicators rather than the thousands that are currently recommended. Priority health indicators should be selected on the basis of public-health importance and the existence of a pragmatic measurement strategy. I believe that focusing attention within and across countries on a restricted set of priority indicators will help show how measurement can foster health improvement. Analyses based on these priority indicators can foster demand from decision makers, the media, civil society, the technical community, and the public for better health measurement. Efforts to strengthen platforms for primary data collection driven by this demand are more likely to gain broad support than are calls for strengthening these platforms that are not linked to specific analytical uses.

The measurement strategy should be designed to produce valid, reliable, and comparable information. If the results of measurement are not comparable over time and across places, they will not be useful for monitoring or assessment. Comparability has not been given

sufficient attention in the development of measurement strategies. Furthermore, every effort should be made to produce statistics for the true quantity of interest instead of proxies. Proxy indicators should be mapped into the quantity of interest and the uncertainty in that mapping quantified.

Crude statistics from primary data collection platforms should be reported as a resource for the analysis and research community, but should be clearly distinguished from corrected statistics for interpretive purposes. The most detailed information possible, including metadata and wherever possible microdata, should be in the public domain.

Statistics used for monitoring and assessment should always be corrected for known biases, and the basis for these corrections should be in the public domain. Predicted statistics should in general not be used for actual monitoring of national or local progress, and should never be used for assessing what works and what does not. Predicted statistics have an important and useful role in helping to inform planning, strategic decision making, and research and development prioritisation when corrected statistics are unavailable. Whenever predicted statistics are used, however, efforts should be made to adequately characterise the uncertainty in the predictions from all sources so that they are interpreted with caution.

All statistics produced for all purposes should have a well documented, explicit data audit trail, which allows a sceptical scientist to entirely replicate the generation of the corrected or predicted statistics. Due attention should be given to protection of public interest in transparency and accountability that derives from having data in the public domain.

The MDG health-related indicators have been used to draw attention to many of the points in this paper. Overall, the set of indicators, the measurement strategies, and the implementation of the MDG health-related indicators is very poor. At least half the indicators are not of high enough public-health importance to warrant major international attention or investments in their assessment. Data availability for developing countries overall is 15%, and even in 2000 when a special effort was made, it reached 44%. Most of these figures are predicted statistics that do not provide a reasonable basis for monitoring progress. The dilemma is that data systems and measurements are so weak that only undernutrition, child mortality, measles immunisation, and attended deliveries have effective measurement strategies in place. WHO, UNICEF, UNAIDS, and other agencies responsible for monitoring the MDGs rely on using predicted statistics for most other indicators so that attention on the MDGs does not fade. Predicted statistics are a reasonable approach for identification of difficulties and for stimulation of policy interest. However, the worldwide community is failing in implementation of viable measurement strategies for real monitoring.

Conflict of interest statement

I was former Executive Director of the Evidence and Information for Policy Cluster at the World Health Organisation. I have no conflicts of interest that may have influenced this work or the conclusions of the manuscript.

References

1 The International Bank for Reconstruction and Development/The World Bank. Global Monitoring Report 2004: policies and actions for achieving the Millennium Development Goals and related outcomes. Washington: The International Bank for Reconstruction and Development/The World Bank, 2004.

2 European Commission. EU report on millennium development goals 2000–2004. Brussels: European Commission, 2005.

3 Healthcare Commission. Our progress: one year on. Annual report 2004/2005. London: Healthcare Commission, 2005. http://www.healthcarecommission.org.uk/_db/_documents/Annual_report_revised.pdf (accessed Feb 21, 2007).

4 Naik G. WHO misses AIDS-treatment goal. *The Wall Street Journal (New York)*, July 12, 2004. http://ww2.aegis.com/news/wsj/2004/WJ040717.html (accessed Feb 21, 2007).

5 Kantrowitz B, Wingert P. More questions than answers: recent results from a major study have challenged many assumptions about women's health. A look at how the trials got started and what to expect next. *Newsweek (New York)*, Feb 28, 2006.

6 Sturcke J and agencies. China pledges transparency over bird flu. *Guardian Unlimited*, Nov 1, 2005. http://www.guardian.co.uk/birdflu/story/0,14207,1606149,00.html (accessed Jan 30, 2007).

7 Boerma JT, Stansfield SK. Health statistics now: are we making the right investments? *Lancet* 2007; **369:** 779–86.

8 WHO. Indicators for monitoring national drug policies. Geneva: World Health Organization, 1999.

9 WHO. National AIDS programmes: a guide to indicators for monitoring and evaluating national antiretroviral programmes. Geneva: World Health Organization, 2005.

10 UNAIDS/WHO/UNICEF/UNFPA/USAID/UNESCO/World Bank/Measure DHS/Family Health International. National AIDS programmes: a guide to indicators for monitoring and evaluating national HIV/AIDS prevention programmes for young people. Geneva: World Health Organization, 2004.

11 World Health Organization. Health metrics network. http://www.who.int/healthmetrics/ (accessed Jan 30, 2007).

12 United Nations. Millennium indicators database. United Nations, 2006. http://unstats.un.org/unsd/mi/mi_goals.asp (accessed Jan 30, 2007).

13 Shengelia B, Tandon A, Adams OB, Murray CJL. Access, utilization, quality, and effective coverage: an integrated conceptual framework and measurement strategy. *Soc Sci Med* 2005; **61:** 97–109.

14 Murray CJL, Tandon A, Salomon JA, Mathers CD, Sadana R. Cross-population comparability of evidence for health policy. In: Murray CJL, Evans DB, eds. Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization, 2003: 705–13.

15 Gwatkin DR. How much would poor people gain from faster progress towards the Millennium Development Goals for health? *Lancet* 2005; **365:** 813–17.

16 Gakidou EE, Murray CJL, Frenk J. Defining and measuring health inequality: an approach based on the distribution of health expectancy. *Bull World Health Organ* 2000; **78:** 42–54.

17 Gakidou E, King G. Measuring total health inequality: adding individual variation to group-level differences. *Int J Equity Health* 2002; **1:** 3

18 Ustun TB, Chatterji S, Mechbal A, Murray CJL, WHS collaborating groups. The world health surveys. In: Murray CJL, Evans DB, eds. Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization, 2003: 797–808.

19 Walker N, Grassly NC, Garnett GP, Stanecki KA, Ghys PD. Estimating the global burden of HIV/AIDS: what do we really know about the HIV pandemic? *Lancet* 2004, **363:** 2180–85.

20 Ezzati M, Martin H, Skjold S, Vander Hoorn S, Murray CJL. Trends in national and state-level obesity in the USA after correction for self-report bias: analysis of health surveys. *J R Soc Med* 2006; **99:** 250–57.

21 Murray CJL, Kulkarni SC, Ezzati M. Understanding the coronary heart disease versus total cardiovascular mortality paradox: a method to enhance the comparability of cardiovascular death statistics in the United States. *Circulation* 2006; **113:** 2071–81.

22 Lloyd-Jones DM, Martin DO, Larson MG, et al. Accuracy of death certificates for coding coronary heart disease as the cause of death. *Ann Intern Med* 1998; **129:** 1020–026.

23 Engel LW, Strauchen JA, Chiazze L, Jr. et al. Accuracy of death certification in an autopsied population with specific attention to malignant neoplasms and vascular diseases. *Am J Epidemiol* 1980; **111:** 99–112.

24 Mapping Malaria Risk in Africa (MARA). MARA-LITe database. MARA/ARMA, 2004. http://www.mara.org.za/ (accessed Jan 30, 2007).

25 United Nations Population Division. QFIVE: microcomputer program for child mortality estimation. New York: United Nations, 1989.

26 United Nations Population Division. Manual X: indirect demographic estimation. New York: United Nations, 1983.

27 Hill K, Choi Y, Timaeus I. Unconventional approaches to mortality estimation. *Demogr Res* 2005; **13:** 281–300.

28 UNAIDS. Improving estimates and projections of HIV/AIDS. Report of a meeting of the UNAIDS/WHO reference group for "Estimates, Modelling and Projections". San Lorenzo de El Escorial, Spain, Dec 4–6, 2002. Geneva: UNAIDS/WHO, 2003

29 Bryce J, Boschi-Pinto C, Shibuya K, Black RE, WHO Child Health Epidemiology Reference Group. WHO estimates of the causes of death in children. *Lancet* 2005; **365:** 1147–52.

30 Lawn JE, Wilczynska-Kotende K, Cousens S. Estimating the causes of 4 million neonatal deaths in the year 2000. *Int J Epidemiol* 2006; **35:** 706–18.

31 Rowe AK, Rowe SW, Snow RW, et al. The burden of malaria mortality among African children in the year 2000. *Int J Epidemiol* 2006; **35:** 691–704.

32 Razum O, Gerhardus A. Methodological triangulation in public health research—advancement or mirage? *Trop Med Int Health* 1999; **4:** 243–44.

33 UNICEF. Child Mortality Database. http://newsite.unicef.org/childmortality/bestestimate.php?areaidx=21 (accessed Feb 28, 2007).

34 Murray CJL, Lopez AD. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Cambridge, Harvard University Press on behalf of the World Health Organization and the World Bank, 1996.

35 Murray CJL, Lopez A. Quantifying global mortality, disability, and the contribution of risk factors: results of the Global Burden of Disease Study. *Lancet* 1997; **349:** 1436–42.

36 WHO. Communicable disease global atlas. Geneva: World Health Organization, 2005. http://www.who.int/globalatlas (accessed Jan 30, 2007).

37 UN. World population prospects. The 2004 revision. New York: United Nations, 2005.

38 WHO. The world health report 2000—health systems: improving performance. Geneva: World Health Organization, 2000.

39 Sobel DL, Davis Noll BA, Fernandez Bogado B, TCC Group, Price ME. The FederalInstitute for Access to Public Information in Mexico and a Culture of Transparency. Project for Global Communication Studies, Annenberg School for Communication,University of Pennsylvania. A Report for the William and Flora Hewlett Foundation, February 2006. http://www.freedominfo.org/documents/mex_report_fiai06_english.pdf (accessed Jan 30, 2007).

40 BBC News. Indians win right to information. BBC News (London), Oct 12, 2005. http://news.bbc.co.uk/1/hi/world/south_asia/4334080.stm (accessed Feb 21, 2007).

41 International Freedom of Expression eXchange. Update. Alliance 72 submits its recommendations on the "Transparency Law" to Congress. International Freedom of Expression eXchange, Feb 8, 2006. http://www.ifex.org/en/content/view/full/72127 (accessed Feb 21, 2007).

42 Dawn. Pakistan: changes to freedom of information law sought. Centre for Peace and Development Initiative calls for lesser fees to request information. Asia Media, Feb 14, 2006. http://www.asiamedia.ucla.edu/article.asp?parentid=39190 (accessed Feb 21, 2007).

43 IOM. Protecting data privacy in health services research. Washington DC: Institute of Medicine, 2000.