Evaluating national cause-of-death statistics: principles and application to the case of China
Chalapati Rao,1 Alan D. Lopez,1 Gonghuan Yang,2 Stephen Begg,1 & Jiemin Ma2

Abstract Mortality statistics systems provide basic information on the levels and causes of mortality in populations. Only a third of the world’s countries have complete civil registration systems that yield adequate cause-specific mortality data for health policymaking and monitoring. This paper describes the development of a set of criteria for evaluating the quality of national mortality statistics and applies them to China as an example. The criteria cover a range of structural, statistical and technical aspects of national mortality data. Little is known about cause-of-death data in China, which is home to roughly one-fifth of the world’s population. These criteria were used to evaluate the utility of data from two mortality statistics systems in use in China, namely the Ministry of Health-Vital Registration (MOH-VR) system and the Disease Surveillance Point (DSP) system. We concluded that mortality registration was incomplete in both. No statistics were available for geographical subdivisions of the country to inform resource allocation or for the monitoring of health programmes. Compilation and publication of statistics is irregular in the case of the DSP, and they are not made publicly available at all by the MOH-VR. More research is required to measure the content validity of cause-of-death attribution in the two systems, especially due to the use of verbal autopsy methods in rural areas. This framework of criteria-based evaluation is recommended for the evaluation of national mortality data in developing countries to determine their utility and to guide efforts to improve their value for guiding policy.

Keywords Mortality; Cause of death; Statistics/standards; Data collection/standards; Models, Statistical; Evaluation studies; China (source: MeSH, NLM).

Mots clés Mortalité; Cause décès; Statistique/normes; Collecte données/normes; Modèle statistique; Etude évaluation; Chine (source: MeSH, INSERM).

Palabras clave Mortalidad; Causa de muerte; Estadística/normas; Recolección de datos/normas; Modelos estadísticos; Estudios de evaluación; China (fuente: DeCS, BIREME).

Introduction Cause-of-death statistics are basic outputs of health information systems and are fundamental to health development strategies. Civil registration systems are the main source of such information and are most useful when both coverage and medical certification are high.

Although many countries invest considerable resources in the establishment and maintenance of systems to monitor the levels, patterns and causes of mortality, currently there is no accepted framework for assessing the quality of this information. Several factors are likely to influence the usability of such information. Ruzicka & Lopez (1) initially proposed five criteria for evaluating quality with respect to data reported to WHO. Subsequently, Mahapatra & Rao (2) applied these criteria to mortality registration systems in India, and proposed four additional components for assessing the usability of cause-of-death information to guide health policy.

The aim of this paper is to position these criteria within an overall framework that others might find useful when evaluating cause-of-death information. The application of this framework is illustrated using China as an example, in part because there are competing systems for reporting mortality statistics in that country, but also because information about the quality of data in China is of interest for global descriptive analyses of health conditions.

For each criterion, a subjective rating of data quality was made using a three-tiered scale:
1. satisfactory;
2. unsatisfactory; or
3. unknown.

Note that the ratings assigned reflect the judgement of the authors, based on a review of available evidence for the indicators. It should be noted that these criteria are intended to assist in the evaluation of data emanating from vital registration systems, and not of the systems themselves. System evaluation requires a wholly different set of criteria related to staffing, operational procedures, efficiency and comprehension by government and...
policy-makers (3). Although these are important issues, they are beyond the scope of this paper.

**Conceptual framework**

Given that cause-of-death statistics are used to guide important decisions regarding health policy and research, it is essential that the criteria for evaluating their utility are able to identify potential sources of misinformation. Identification of errors and biases in the data will increase confidence in using the statistics. Similarly, knowledge that data-collection methods are consistent across time and place (for subnational analyses) is essential to assess policy utility. These and related considerations provide the basis of the framework we propose.

A useful starting point is to focus on **generalizability**, i.e. how representative are the statistics with respect to the population(s) of interest? We propose two criteria to assess generalizability.

- **Coverage.** To what proportion(s) or sectors of the population do the data refer? Coverage may include urban/rural, geographical or administrative sectors, or coverage of specific socioeconomic groups. Coverage is also relevant in the context of sample-based systems.
- **Completeness.** Within the population covered by the systems for recording cause of death (e.g. urban areas only), what percentage of deaths are registered with a cause?

Another consideration is **reliability**, i.e. how consistent are the data with respect to epidemiological expectations. For example, previous research suggests that proportionate mortality from various causes changes in a predictable fashion over time as overall mortality declines with socioeconomic development and disease control (4–6). Equally, large annual fluctuations in levels of mortality from such causes as cancer or stroke would be unexpected. We propose two criteria under this rubric:

- **consistency of cause patterns with general levels of mortality;**
- **consistency of cause-specific mortality rates over time (e.g. over 5 years).**

Arguably, the most important elements of any framework for evaluating data quality are indicators of the **validity** of the data. For this, we propose three criteria:

- **Content validity:** that is, do the statistics actually show what they purport to show? Is there information on the proportion of deaths certified by medical practitioners? Are procedures for certification and coding comparable across the population covered by the statistics? Are there periodic comparisons of the accuracy of diagnoses with some “gold standard”? Information on such issues allows the user to assess the validity of the data.
- **Use of ill-defined categories and codes:** what proportion of deaths is assigned an ill-defined cause, or otherwise poorly diagnosed (e.g. cancer without mention of primary site; injury that has not been classified according to whether or not it was intentional. Such diagnoses are of little value for policy or epidemiological research.
- **Incorrect or improbable age or sex dependency:** several causes of death are strongly age-dependent (e.g. stroke or perinatal causes) or uniquely sex-specific (e.g. prostate cancer and maternal haemorrhage). Departures from these patterns reflect serious concerns with the validity of the data.

The final element of the framework groups together criteria to assess the **policy relevance** of the data. We propose two criteria, both of which are of critical importance for policy-makers.

- **Timeliness:** data on causes of death that are more than, say 2 years out of date rapidly lose their relevance for policy and programme purposes.
- **Geographical disaggregation:** statistics that identify the extent and pattern of health differentials in populations within health service jurisdictions are critical for informing policies to reduce health inequalities and target interventions.

**Implementation**

Ease of application is the key to the success of an evaluation framework. In most cases, the criteria we propose can be evaluated directly from the cause-of-death statistics, administrative information on the registration system or other information that is readily available. For example, completeness can be calculated using indirect demographic methods (7) and available estimates of total mortality (8, 9). Alternatively, direct methods based on the dual record system (10) could be employed by conducting independent re-surveys. Some of the indicators of data reliability and validity can be calculated directly from the data, such as consistency over time, age or sex dependency, and proportions of deaths assigned ill-defined codes. Timeliness is readily apparent, and the availability of geographical statistics can be determined from data management systems in place. The only criterion for which assessment might prove difficult is content validity, as this requires either supplementary tabulations (e.g. proportion of deaths medically certified, or extent of lay reporting), or reference to the research literature on formal validation studies (11–17).

Another critical issue is the standards by which quality will be judged. Although a detailed scoring system for each criterion would certainly increase the discriminatory power of the framework, in most cases it is likely to be sufficient to judge performance as simply “satisfactory” or “unsatisfactory”, or if insufficient information is available, “unknown”. The thresholds we propose for differentiating the first two of these categories are based on a common-sense understanding of which data are usable and which are not. Coverage of less than the entire population is obviously unsatisfactory, unless based on representative sampling. Completeness levels of less than 90–95%
could imply biases in the data towards conditions more commonly treated in health facilities, or medico-legal cases (for which a police or judicial enquiry is required to ascertain the cause of death). Consistency of cause-of-death data with overall mortality might be assessed in comparison with cause-of-death models, using the criterion that more than two standard deviations from predicted levels reflects unsatisfactory performance (18), unless there is convincing evidence to suggest that the observed deviations are likely to be real.

Similarly, annual fluctuations in death rates from leading causes that are greater than about 2–3% suggest quality concerns and an “unsatisfactory” rating. Because it is difficult to assess content validity objectively, it may be captured by other criteria. For example, a high proportion of ill-defined deaths (> 10%) suggests unsatisfactory certification and coding procedures, and implies inadequate medical input into the certification process, leading to an “unsatisfactory” rating for content validity. Any indication that causes of death are being assigned to impossible or improbable age or sex categories for specific diseases based on known epidemiological considerations would warrant caution, and hence an “unsatisfactory” score. Finally, given the need for up-to-date information, statistics that are more than two calendar years old are likely to be of limited value.

Evaluation of Chinese mortality statistics systems

In this section, we apply the criteria listed above to the two mortality data systems currently in operation in China: the Ministry of Health-Vital Registration (MOH-VR) system, and the Disease Surveillance Points (DSP) system. A detailed description of the history and organization of mortality registration in China is described elsewhere (19). In urban areas, certification of cause of non-domiciliary deaths is done by the hospital where the deceased was treated. In rural areas both systems employ “verbal autopsies” (VA) to ascertain the cause of death, supplemented by information from medical records kept on the household. Verbal autopsy is a process that involves a structured interview with relatives of the deceased to collect information on symptoms and events during the illness that preceded death. This information is subsequently reviewed to determine the probable cause of death.

Age-standardized mortality rates differ between the two systems for deaths due to communicable diseases, noncommunicable diseases and external causes (19). Protein–energy malnutrition and meningitis are among the leading causes of childhood mortality according to the DSP, but are absent from the 10 leading causes in the MOH-VR. Among the elderly, there are important differences in the reporting of proportions of deaths from ischaemic heart disease (7.7% in DSP; 10.5% in MOH-VR), and lung cancer (4.3% in DSP; 5.7% in MOH-VR). These observations lead to uncertainty about the true level and pattern of cause-specific mortality in China. In the following sections, we systematically assess the operational characteristics of the two systems using the nine specific criteria described above to illustrate how they might be applied in practice to mortality registration data.

Coverage

The MOH-VR covers about 8% (110 million) of the national population, whereas the DSP covers about 1%. The sample sites in the DSP system remain constant over time, and are representative of the national population (satisfactory). In contrast, the sites represented in the MOH-VR data are not representative because they are focused almost entirely in the eastern region (19). Also, the number of sites represented in the MOH-VR data varies from year to year, and there are no specific criteria for inclusion or exclusion. During the period 1988–2000, annual fluctuations in the reported sample population and registered deaths exceeded 2% compared to the previous year in eight instances (Fig. 1). These fluctuations raise doubts about the validity of the statistics generated by this system (unsatisfactory).

Completeness

To assess completeness, the DSP conducts independent surveys on a sample of about 5000 households in each province once every 3 years. The results from the 1998 survey (Table 1) suggest that significant under-reporting occurs at all ages, particularly in children. Uncertainty exists about the independence of the two data collection efforts (an underlying principle in the application of these statistical methods (10)) because the village doctors could be involved in both systems. However, an indirect demographic assessment of completeness (the growth-balance method (20)) also suggests under-reporting of adult mortality (30%) in the DSP system (9); hence we rate it as unsatisfactory on this criterion. (The Brass Growth-Balance technique assesses the completeness of death registration at ages above 5 years by comparing the age-distribution of reported deaths with the expected age distribution based on the observed population age–sex structure, assuming the population is closed to migration.)
Chalapati Rao et al.

No data are available for direct assessment of completeness in the MOH-VR system. Conventional life-table analysis of observed age-specific mortality rates from this system yields implausibly low levels of adult mortality (9). This was confirmed by comparing observed and model-predicted levels of adult mortality from the modified logit model life table system (21) (Fig. 2). The system uses levels of child mortality ($a_{0}$) to predict adult mortality ($a_{15}$), based on regression equations developed from observed relationships between these two parameters in empirical data. We used the model to predict a $a_{15}$ for every observed $a_{0}$ for males in the MOH-VR data series from 1989–2000. Observed values of $a_{15}$ are consistently and markedly lower than the values predicted by the model. Similar results were obtained for females. These analyses suggest that levels of adult mortality as measured from MOH-VR data are implausibly low, hence we rate the MOH-VR system as unsatisfactory on this criterion.

Content validity
To assess content validity, we focused on the specific collection mechanisms for collecting data on individuals who had died, and the process used to compile statistics based on underlying causes of death. Following new legislation in 1992 (22), a uniform death certification procedure has been practised for deaths in health facilities in China, using the international form of the medical certificate of cause of death (23). For deaths that occur in the home, both systems rely on information obtained from family members, with the DSP using verbal autopsy interviews, whereas the MOH-VR sometimes records family opinion alone, without detailed investigation. The DSP carries out coding and classification of underlying causes at the central level, which reduces inter-rater errors (i.e. differences between different coders) in these procedures, whereas the MOH-VR conducts coding at the provincial level.

The final statistical tabulations from both systems include all data, without differentiation of institutional deaths from home deaths, and home deaths are not differentiated according to whether or not a medical opinion was obtained as to the cause. The validity of the content of data sets from both systems is therefore questionable, given the absence of these supplementary statistics. No studies have been published that measure the accuracy of death certificates in comparison with clinical records (unsatisfactory).

Use of ill-defined categories
In any population, there are invariably some deaths for which the cause is difficult to ascertain. These are coded according to chapter XVI in the International Classification of Diseases (ICD)-9 (24). Few deaths were assigned codes in this category over the period 1995–99 (average of 2.5 % in DSP; 5% in the MOH-VR). This suggests that when deaths are registered, the cause is determined with a high degree of specificity. Hence, we rate both systems as “satisfactory” under this criterion.

Other categories that could also be assessed under this criterion are septicaemia (A41 in ICD-10), cardiac arrest (I46), heart failure (I50) injuries of undetermined intent (Y10–Y34), and malignant neoplasms of ill-defined, secondary and unspecified sites (C76–C80).
Age and sex dependency
Certain causes are uniquely age-specific (e.g. perinatal conditions) or sex-specific (e.g. prostate cancer), and although misclassifications of this nature could appear in the final statistics as a result of clerical error, adequate screening of aggregated data with subsequent review should correct them. Data from neither system during 1995–99 revealed any such instances. We therefore rate both data sources “satisfactory” under this criterion.

Furthermore, certain diseases and conditions mainly cause mortality in specific age groups (e.g. diarrhoeal diseases or meningitis in children, road traffic accidents in young adults, chronic obstructive pulmonary disease (COPD) and stroke in the elderly). Deviations from expected age patterns of mortality for specific conditions, unless explained by local epidemiological phenomena, raise doubts about the quality of certification and coding. Fig. 3 shows age patterns of numbers of deaths due to four selected causes from the DSP data. All appear plausible and consistent with data from countries with high-quality mortality statistics systems (25). Age patterns for 20 leading causes from each system were examined, and found to be similarly consistent, and hence we rate both data sources “satisfactory” on this criterion.

Consistency with general mortality levels
In the absence of statistics from complete civil registration systems, cause of death models are used to predict the cause structure of mortality by age and sex, for a given level of all-cause mortality (6). We compared observed and model predicted-proportionate composition of mortality by age for males and females according to three broad groups of causes of death. Fig. 4 shows the results for data from the DSP system pooled for the three years 1997–99. For communicable and noncommunicable diseases, observed proportions for age–sex–cause combinations were within two standard deviations of the predicted value (Fig. 5). However, the observed proportions for injuries were significantly higher than the predicted value, especially among females. Although the models do not include Chinese data, the weight of historical evidence raises uncertainty about the validity of these deviations (18). Medico-legal considerations resulting in more complete registration of deaths from injury could possibly explain the findings. Detailed studies that assess the causes of deaths missed by registration could provide further insight. Evaluation of data from the MOH-VR for 1999 yielded similar results. In view of this uncertainty, we rate both systems as “unknown” under this criterion.

Consistency over time
Apart from changes resulting from new diagnostic methods, or revisions to the classification of diseases (ICD), one would expect a uniform trend in cause-specific mortality in the absence of any disease outbreak or natural disaster. We examined the time trend in cause-specific mortality proportions from 10 leading causes in males and females from each of the systems during 1997–99. Poisson regression methods were applied to calculate 95% confidence intervals for each annual cause-specific mortality proportion. A significant fluctuation in the mortality proportion was observed in only one case (COPD) in the DSP system (satisfactory).

Timeliness
The compilation of statistics including data cleaning (i.e. the identification and correction of errors in the data that could result from incorrect data-entry or processing) and mechanisms of re-enquiry to obtain information pertaining to a potential data error is time consuming. On the other hand, it is technically incorrect and politically unhelpful to publish questionable data. A time lag of about 2 years may be considered reasonable for the release of final tabulations. Statistics from the DSP were available in a timely manner during the 1990s. However, there has been no public release of statistics since 1999. Data from the MOH-VR system are available only for internal use within the Ministry of Health. There is no published analysis of these data within China, and basic statistics are released only to the World Health Organization (unsatisfactory).
Usability for subnational purposes

Both the MOH-VR and the DSP systems produce statistics according to a socioeconomic stratification of rural counties. Life tables constructed from the two systems show a plausible decline in life expectancy and rise in risks of child and adult mortality with decreasing socioeconomic status (see Table 2).

These socioeconomic strata represent subpopulations dispersed throughout the country, not by geographically determined health administration districts, yet health service provision and responsibility is generally organized according to geographical entities. Absence of these data from both systems precludes such analysis. Moreover, the DSP system was designed for health policy and monitoring at the national level and the sample sizes are too small for subnational assessment, even if pooled over several years.

The MOH-VR data are also constrained in this respect, in that they are primarily limited to the urban population living in the eastern China. Hence, we rate both data sources “unsatisfactory” under this criterion.

Discussion

Many countries invest substantial resources in the development and maintenance of data systems for recording cause of death. We have developed a set of criteria which we believe will assist assessments of cause-of-death information from such systems. They are relatively easy to apply, with most being measurable from the data themselves. The criteria are not intended as a tool for evaluating the registration system itself, however, although one might argue that data quality is a reasonable proxy for the functioning of the system that generates the data. Some criteria (e.g. validity) require additional background information or research results, which may not be readily available. For some purposes, the criteria might not possess adequate discriminatory power to distinguish between the quality of different cause-of-death data sets.

The application of the criteria to China suggests that there is an urgent need to address various systemic issues, particularly completeness in the DSP system. Although the system was able to collect data on about 85% of deaths in the late 1990s, higher levels of completeness are necessary to obtain largely unbiased statistics on cause-specific mortality.

The content validity of cause-of-death reporting in the registration system needs further research. Possible biases in the observed data due to incomplete registration remain unidentified. Additional biases may arise due to the specific sociocultural context within which the system for recording cause of death operates; a better understanding of this will help in interpreting the data. A good example is the larger than predicted proportion of deaths from injury among young women in China (Fig. 4). Furthermore, for individual deaths, there is currently no measure of the accuracy of cause attribution or of the subsequent ICD coding of the underlying cause.

The use of verbal autopsy methods in the DSP raises further concerns. Although these are necessary, due to the large proportion of deaths that occur in the home, a detailed validation of the methods used in China is required. Assessments of
validity and reliability in both urban and rural areas of China are being implemented to guide the integration of the two systems into a unified nationally representative system for mortality statistics.

This evaluation has suggested areas where investment is urgently needed to improve data quality. Other countries where there is uncertainty about data quality may wish to evaluate their statistics against these criteria and conduct similar research.

Periodically carrying out such assessments will greatly enhance the utility of cause-of-death data.

Acknowledgements
This work was sponsored by the National Institute of Aging Grant 1 – PO1-AG17625.

Competing interests: none declared.
References


