METHODOLOGY

Validation of verbal autopsy procedures for adult deaths in China

Gonghuan Yang\(^1\), Chalapati Rao\(^2\), Jiemin Ma\(^1\), Lijun Wang\(^3\), Xia Wan\(^1\), Guillermo Dubrovsky\(^2\) and Alan D Lopez\(^2\)*

Accepted 8 August 2005

**Background**

Vital registration of causes of death in China is incomplete with poor coverage of medical certification. Information on the leading causes of mortality will continue to rely on verbal autopsy (VA) methods. A new international VA form is being considered for data collection in China, but it first needs to be validated to determine its operating characteristics.

**Methods**

Detailed medical records and clinical evidence for 3290 deaths (mostly adults) among residents of six cities representative of the urban Chinese population were reviewed by a panel of physicians and coded by experts to establish a reference underlying cause of death. Independently, families of the deceased were interviewed using a structured symptomatic questionnaire and a separate death certificate was prepared for each matching case (2102). Validity of the VA procedure was assessed using standard measurement criteria of sensitivity, specificity, and positive predictive value.

**Results**

VA methods perform reasonably well in identifying deaths from several leading causes of adult deaths including stroke, several major cancer sites (lung, liver, stomach, oesophagus, and colorectal), and transport accidents. Sensitivity was less satisfactory in detecting deaths from several causes of major public health concern in China including ischaemic heart disease, chronic obstructive pulmonary disease, diabetes, and tuberculosis, and was particularly poor in diagnosing deaths from viral hepatitis, hypertension, and kidney diseases.

**Conclusions**

VA is an imprecise tool for detecting leading causes of death among adults. However, much of the misclassification generally occurs within broad cause groups (e.g. CVD, respiratory diseases, and liver diseases). Moreover, compensating patterns of misclassification would appear to suggest that, in urban China at least, the method yields population-level cause-specific estimates that are reasonably reliable. These results suggest the possible utility of these methods in rural China, to back up the low coverage of medical certification of cause of death owing to poor access to health facilities there.

**Keywords**

Verbal autopsy, validation, China, disease surveillance point (DSP) system, mortality, vital registration

---

1 Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, 5, Dong Dan San Tiao, 100005 Beijing, China.
2 School of Population Health, University of Queensland, Public Health Building, Herston Road, Herston, Qld 4006, Australia.
3 National Centre for Disease Control and Prevention, 27 Nanwei Road, 100050 Beijing, China.
* Corresponding author. E-mail: a.lopez@sph.uq.edu.au

---

**Introduction**

Ascertaining the cause of a death in the absence of medical attention is a difficult task, especially in the absence of medical records for the deceased. The only recourse then lies in the use of ‘verbal autopsy’ (VA) interviews to gather information about symptoms and events during the period leading up to death, which is then reviewed to derive a probable cause. The VA methodology has been used in several developing countries to
collect information on cause-specific mortality, given the basic necessity of such data for health policy and planning. These methods have been previously used to investigate epidemics and for specific intervention evaluation. They have also been implemented in national mortality surveillance systems, principally in India, China, and Tanzania.

Recall bias leads to uncertainty in the diagnosis of cause of death from such retrospective interviews. Other issues that could influence the quality of VA data include questionnaire design, choice of interviewers and respondents, cause of death ascertainment mechanisms, and procedures for coding and tabulation of the data. Implementation of standardized practices could minimize bias that results from variations in these characteristics. However, there is a critical need to measure the validity of such methods in correctly ascertaining the leading causes of death and others of epidemiological significance in specific populations, if there is to be any confidence in the data collected through VAs.

By definition, validation involves a comparison of the underlying cause of death derived from the VA with a ‘true’ or reference underlying cause for the same death, derived from either a pathological autopsy (‘gold standard’) or clinical records (next best alternative). Non-availability of a reference diagnosis for the underlying cause is an elementary constraint in measuring the validity of VA procedures. This effectively excludes validation in rural areas of China where few deaths have appropriate clinical records. Causes of childhood death have been the focus of several VA validation studies in Africa and South Asia, where child mortality remains a major public health problem. Disease/condition-specific algorithms have been developed and validated in these studies, providing a wealth of empirical evidence on the usefulness of VA for childhood mortality.

Little work, however, has been done on the validation of VA for adult deaths. What has been done is largely limited to the investigation of deaths due to HIV and maternal causes, although VA methods have also been used to correct for ill-defined causes in India and Thailand, and to evaluate the reliability of routine vital registration in Jordan. Populations in developing countries are faced with an increasing burden from early adult mortality, from infections (HIV and tuberculosis), non-communicable diseases (cancers, cardiovascular disease, and diabetes), and injuries. In contrast to the major causes of child mortality, some of these conditions are difficult to diagnose even in clinical settings, let alone using VA, owing to complex symptomatology. Nevertheless, the use of VA is unavoidable in many populations with poor access to health facilities and with inefficient cause of death certification and registration mechanisms.

To establish the public health utility of VA derived data, its validity first needs to be assessed. In this study, we validate a set of VA procedures that have been proposed for use in the disease surveillance point (DSP) system in China. A detailed description of the DSP is available elsewhere. In this system, retrospective interviews are conducted among relatives to ascertain the cause of death for domiciliary deaths occurring in a nationally representative sample of population clusters distributed throughout China. The VA methods used to date have varied across provinces and counties. The specific aims of this study were to ascertain the validity of a standard set of VA procedures developed by an international collaboration to identify leading causes of adult deaths in China, and second, to identify patterns of misclassification error for different causes, and the reasons for such misclassification (Setel P, Ras C, Hemed Y et al., submitted for publication). Knowledge of such misclassification patterns, if broadly applicable, can be used to correct data from routine death reporting systems to more reliably estimate overall cause-specific mortality patterns in China. Not only will this better inform national disease prevention and control strategies, but it is also of obvious interest for assessing global mortality patterns and burden of disease.

Materials and methods

The study was conducted in six cities representative of the urban population residing in different geographical regions of China, namely Beijing (North), Ha'erbin (North-east), Shanghai (East), Guangzhou (South), Wuhan (Central), and Chengdu (South-west) (Figure 1).

The VA validation study formed one component of a broader study to evaluate the quality of mortality statistics from the routine death registration system in China. Deaths were selected into the study based on the frequency of occurrence of causes reported in the routine system, but also included adequate numbers of certain less-frequent conditions likely to be of local epidemiological significance (e.g. hepatitis). In order to ensure that comparative data (from medical records and VA) were available for 20 leading causes or causes otherwise of interest, and assuming a failure rate of about one-third (based on pilot investigation), 3300 cases were selected, resulting in just over 2100 VAs. The list of causes chosen for validation can be seen from Table 2. Certain high frequency causes of death such as ischaemic heart disease (IHD), cerebrovascular disease, and chronic obstructive pulmonary disease (COPD) were under-sampled, in order to achieve adequate samples (>25 deaths) of less-frequent causes of interest for validation, such as site-specific cancers, renal disorders, and infectious diseases.

The criteria for selection into the study were as follows:

(i) The deceased should have been a resident of the city.
(ii) The death should have occurred between June 2002 and November 2002.
(iii) The death should have occurred in a tertiary care health facility with adequate diagnostic facilities.
(iv) The medical records for that death should be available for determining cause of death.

Altogether, ~100 hospitals in the six cities provided the primary records for this study. An overview of the data collection, processing, and analysis protocol for validation of the VA procedures is shown in Figure 2. Roughly equal numbers of deaths (500–600) were selected in each of the six study sites, yielding a total of 3290 deaths. For each selected death, an international death certificate was derived, through expert review of the medical records, by a panel of three physicians blind to the cause of death certificate filed at registration. To assist this review, trained personnel abstracted relevant clinical information, diagnostic results, and treatment details from the medical record onto a hospital information sheet. For some cases, photocopies of specific documents from the medical record were attached as supporting evidence. The identity of the deceased was kept confidential. Reference (expert review) death certificates.
were coded by the World Health Organization Collaborating Centre for Classification of Diseases in Beijing to select and code the underlying cause of death according to the core three character code, as recommended by the International Classification of Diseases, Tenth Revision (ICD-10). This underlying cause, based on expert review of the medical records, was used as the reference or ‘gold standard’ diagnosis for assessing validity in this study.
From the original set of 3290 deaths, 363 cases were excluded owing to insufficient evidence being available to derive an underlying cause of death. For the remaining 2927 deaths, the address of the decedent was obtained from the hospital record, and trained field investigators, blind to the cause of death from either the routine system or the expert review, conducted interviews using standard VA instruments.31 These included modules to record demographic characteristics, a narrative of the illness before death as provided by the respondent, a symptom duration checklist, and a section to note any mention of the illness prior to death from health services staff or records. The instruments were pilot tested on a sample of 200 deaths in Beijing, and revised to incorporate Chinese terminology and local perceptions of specific symptoms and signs.

Non-response to the VA interview was 6.5%, mostly in Beijing, Shanghai, and Guangzhou. A further 21% of cases could not be interviewed because of incomplete address or out-migration of the household of the deceased, with Shanghai accounting for about one-third of these cases. Owing to funding constraints, the study was brought to a close on reaching the preliminary target of 2100 deaths, with no further pursuit of losses to follow-up. A comparison of the proportionate distribution of deaths by cause from the original sample (2927 cases with expert diagnoses) and the excluded cases (825 cases not investigated by VA) suggests that the exclusion was non-differential by cause (Figure 3).

Completed VA instruments were submitted for review to an independent panel of two physicians in Beijing, who reached consensus as to the content and ordering of the causes on a standard death certificate for each death. Here too, adequate care was taken to maintain confidentiality. Reviewing physicians applied standard clinical judgment to ascertain causes of death, using diagnostic guidelines similar to those described for use in other settings.32 They also recorded a subjective opinion as to the strength of evidence for each cause. These certificates were also sent to the WHO Collaborating Centre for selection and coding of the underlying cause. Procedures were established by the Centre to ensure that the expert and VA death certificates for the same individual were coded independently.

All information from the routine death certificates, hospital information sheets, expert death certificates, VA questionnaires (separate for neonates, under-fives, and adults), and VA-derived death certificates was entered twice and cleaned in an SAS data management program. Underlying causes from the expert and VA derived death certificates were tabulated at ICD 3 character level, and according to the ICD Mortality Tabulation List 133, as recommended by the WHO. Validation characteristics of VA diagnoses were assessed against categories in this Tabulation List. The parameters measured were sensitivity, specificity, and positive predictive value (PPV), using standard methods, each with 95% confidence intervals. Finally, misclassification was assessed from a matrix of cause attribution of deaths as assigned by the expert review and by VA.

Sensitivity and PPV convey subtly different messages about the ability of VA to identify deaths owing to a particular cause. Sensitivity assesses, from an epidemiological perspective, the probability that the VA correctly diagnoses true deaths from the cause of interest, while PPV, from a clinical perspective, measures the true chance of death from this cause, if diagnosed as such by VA. PPV is dependent on the proportionate distribution of deaths and the observed sensitivity for each cause of interest in the study sample, and this should be borne in mind when extrapolating the findings to make inferences about community-based cause of death patterns.

As a rough guide, VA might be considered to have ‘good’ validity for diseases or conditions for which sensitivity is >75%; for those with sensitivity in the range 50–75%, ‘tolerable’ validity; and when <50%, ‘poor’ validity. These thresholds are admittedly arbitrary, but have some commonsense appeal. For most
purposes, a process that correctly diagnoses the cause of death in <50% of cases is unlikely to be of much public health value.

Results

Table 1 provides an overview of key qualitative characteristics of the VA interviews conducted for this study. Several points are worth noting. First, the study sample essentially comprised adult deaths only, and hence we have excluded the 63 childhood deaths, mostly due to perinatal causes, from the subsequent analyses. About two-thirds of the VA interviews were conducted within a year of the event, which may have affected recall bias. Respondents were first-order relatives (parent, spouse, and offspring) in 85% of cases, and most (81%) were present in the household during the illness preceding death. The majority of households had some information as to the cause of death from contact with health services, and of them, about half could produce some medical documents at interview. Validation was, therefore, largely based on responses to the structured questions on symptoms and duration.

The operational characteristics of the VA procedure are shown in Table 2 for the 20 causes of death selected for this study. The VA procedures we have used appear to have good validity for transport accidents, stroke, several leading sites of cancer (including lung, liver, stomach, oesophagus, and colorectal cancer), and pneumonia. On the other hand, the VA was much less successful in adequately diagnosing several conditions of major public health importance in China such as IHD, COPD, diabetes, tuberculosis, and diseases of the liver, and performed relatively poorly in accurately diagnosing viral hepatitis, hypertension, and kidney diseases.

Table 3 shows the misclassification matrix for leading causes with ‘low’ or ‘tolerable’ sensitivity scores. For each disease, the rows of the matrix indicate the total number of deaths identified as being due to that cause from the VA (VA diagnoses), while the
columns show to which cause each of these deaths was assigned from the review of medical records (MR diagnosis) (i.e. ‘truth’). Thus, for example, IHD caused 206 of the 2102 matched cases in the sample, but only 128 (64%) were correctly identified as such by the VA. Twenty others (10%) were identified by the VA as due to stroke, nine (5%) to hypertensive diseases, and a somewhat smaller fraction to COPD, and a variety of other causes. At the population level, however, the misclassification appears largely compensatory, since an additional 39 deaths were coded to this cause by the VA, but were in reality owing to some other cause, i.e. stroke, hypertensive diseases, COPD, and diabetes.

COPD has been reported as a leading cause of death in China, yet only 59% of COPD deaths in this sample were correctly identified as such by the VA, which misclassified several true COPD deaths as deaths due to stroke (7%) and diabetes (6%). The PPV for COPD was higher (82%), with relatively few strokes and IHD deaths being identified as COPD by the VA. Conversely, despite good sensitivity for pneumonia (75%), the VA misclassified deaths due to COPD (six deaths), stroke, IHD, and several other causes in smaller numbers to this cause, resulting in a poor PPV of 36%. Thus any VA diagnosis of pneumonia among adults in China should prompt a more detailed investigation for a more specific underlying cause.

There is considerable overlap and misclassification of diagnoses between viral hepatitis and chronic liver diseases, which is understandable given the bluntness of VA in differentiating between these conditions. Conversely, cause attribution by the VA for injury deaths was generally very good, as might be expected. The exception was deaths due to falls, for which the sensitivity was 78%, with four cases being misclassified by the VA to cerebrovascular disease.

Discussion
The need for cause of death information to support policy and programme development, and to monitor disease and injury control strategies, is clear. However, given the costs of establishing and maintaining adequate vital registration systems, most developing countries will be dependent for some time on a system of VA to provide information on the main causes of death in their populations, and the way in which these are changing. To be truly useful for public health purposes, the operational characteristics and biases associated with the specific VA procedures in use need to be clearly understood. Validation studies of the type reported here for urban China represent the best and possibly only means to assess confidence in the results of registration systems, which are largely based on VA.

A key element in these studies is the process for determining the reference diagnoses for assessing validity. Diagnoses from medical records are subject to their own information biases, in terms of the nature and strength of clinical, laboratory, or imaging evidence supporting them. Further, there are several nuances in the translation of evidence from medical records into diagnostic labels on the various sections of the death certificate. In this study, we employed a detailed protocol to judge and record the quality of supporting evidence for each case, and applied the principles of cause of death certification prescribed by the ICD.

The chronicity of disease and complex symptomatology makes the identification of a unique underlying cause for each death a particular challenge for many adult deaths. In several instances, despite the presence of the same multiple causes (e.g. COPD and stroke) on the death certificate from both the MR and VA for an individual death, coding experts differed in selection of the underlying cause from each study arm, resulting in a discordant pair. About 10% of all unmatched pairs were owing to such differences in judgement. These problems in the ascertainment of the reference diagnosis could have important implications for measuring the validity of the VA.

For three leading causes of death in China, namely stroke, IHD, and COPD, only 60–70% of deaths identified by the VA in this study sample as owing to these causes actually are. The VA procedures incorrectly assign other diagnoses to these specific diseases, but seemingly equally incorrectly, assign actual deaths from these diseases to other causes. As a result, while the sensitivity of the VA procedure is generally only modest for these causes, the population level estimates of cause-specific mortality
from them is much less distorted, owing to compensatory patterns of misclassification.

Significantly, our results suggest that VA does not perform adequately for several other important causes, including diabetes, hypertensive disease, and diseases of the kidney and liver. For these diseases, symptoms are either not well reported, misunderstood, or cannot be easily differentiated from those leading to other causes. If there is symptom overlap that makes it difficult to distinguish between diseases, the validity of the VA will be affected. Further research is required to identify specific reasons for the poor performance of VA for these diseases.

There is some evidence to suggest that although there might well be poor validity for individual causes, combining causes with common aetiological or risk factor profiles vastly improves sensitivity and PPVs. For instance, if the categories of diabetes, IHD, stroke and hypertensive heart, and renal diseases, having common risk factors (i.e., diet, smoking, hypertension, and obesity) are considered together, the sensitivity for this combined group is ~85%, and PPV also increases. Similar improvements in sensitivity are also observed on combining viral hepatitis and liver diseases. Although, this does not necessarily inspire confidence in VA as a method to correctly identify the true cause of death, it also does not invalidate the method as an approach to estimating population level cause of death patterns, at least for broad causes, within which such compensatory classification typically occurs. These findings would tend to underscore the public health utility of VA, in terms of its efficiency in identifying deaths due to disease clusters, which are principal causes of adult mortality, and thus the target of common public health responses. Further research testing specific questions or algorithms could improve the accuracy of the VA in identifying specific causes, resulting in better individual level cause of death assessment, a matter of importance for specific epidemiological research.

Another factor affecting the validity of VA is the availability of information on the cause of death from the household. In almost 60% of deaths due to cancers, for example, the diagnosis was conveyed to the respondents by health professionals. This information led to an increase in sensitivity for cancers of ~15% compared with households where this information was not known. Similar differences in disease-specific sensitivities were also observed for cerebrovascular diseases, but not for infectious diseases, respiratory disease, or injuries. These findings suggest that contact with health services can substantially affect the information content and hence the validity of VA, at least for some conditions.

Overall, the VA procedures classified only ~4% of all cases in this sample to ill-defined categories, reflecting deaths where either the respondent was not familiar with the deceased’s illness, where there was confusing symptomatology, or perhaps poor interviewing skills. This percentage is low, and is perfectly well for several important causes, or at least disease clusters. Given that 70% or so of China’s population still live in rural areas, innovative methods to assess the accuracy of rural cause of death data need to be rapidly developed as a matter of priority if we are to have more confidence in assessments of national mortality patterns in China, and by implication, the world.

Acknowledgment

This work has been sponsored by the National Institute of Aging Grant 1-PO1-AG17625.

References

Commentary: Verbal autopsy procedure for adult deaths

Vendhan Gajalakshmi1* and Richard Peto2

Estimation of cause of death is more difficult in developing countries because neither health-facility-based information systems nor vital registration provides adequate data on the cause of mortality. In many of the developing countries verbal autopsy (VA) may be a surrogate for death certificates in ascertaining causes of death. VA is a systematic retrospective inquiry of the family members about the circumstances, events, symptoms, and signs of illness prior to death to help determine the underlying cause of death and to classify the broad patterns of mortality. There are two main approaches to conducting a VA.

1 Epidemiological Research Center, Chennai, India.
2 Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), University of Oxford, UK.
* Corresponding author. Epidemiological Research Center, New no. 27, Canal Road, Kilpauk Garden Colony, Chennai 600010, Tamil Nadu, India. E-mail: gajaerc@gmail.com