Verbal Autopsy: Methods in Transition

Edward Fottrell* and Peter Byass

* Correspondence to Dr. Edward Fottrell, Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, SE-901-85 Umeå, Sweden (e-mail: edward.fottrell@epiph.umu.se).

Accepted for publication January 21, 2010.

Understanding of global health and changing morbidity and mortality is limited by inadequate measurement of population health. With fewer than one-third of deaths worldwide being assigned a cause, this long-standing dearth of information, almost exclusively in the world’s poorest countries, hinders understanding of population health and limits opportunities for planning, monitoring, and evaluating interventions. In the absence of routine death registration, verbal autopsy (VA) methods are used to derive probable causes of death. Much effort has been put into refining the approach for specific purposes; however, there has been a lack of harmony regarding such efforts. Subsequently, a variety of methods and principles have been developed, often focusing on a single aspect of VA, and the resulting literature provides an inconsistent picture. By reviewing methodological and conceptual issues in VA, it is evident that VA cannot be reduced to a single one-size-fits-all tool. VA must be contextualized; given the lack of “gold standards,” methodological developments should not be considered in terms of absolute validity but rather in terms of consistency, comparability, and adequacy for the intended purpose. There is an urgent need for clarified thinking about the overall objectives of population-level cause-of-death measurement and harmonized efforts in empirical methodological research.

INTRODUCTION

Cause-of-death data are critical to formulating good public health programs; developing regional, national, and global policies; and implementing and evaluating public health action (1, 2). However, death registration and cause documentation strongly depend on the socioeconomic status of the setting in which death occurs. In less-developed countries, most deaths are neither attended by physicians nor certified medically, so two-thirds to three-quarters of the world’s population remains outside any systematic mortality surveillance (3–7). As an alternative method of documenting cause of death, symptom-based circumstantial inquiries may be used.

Projects in Asia and Africa in the 1950s and 1960s used systematic interviews by physicians to assess causes of death (8). Workers at the Narangwal project in India labeled this technique “verbal autopsy” (VA); the method subsequently spread and developed, particularly during the 1970s, when the World Health Organization suggested lay reporting of health information by people with no medical background (8–11). Today, VA remains the best available approach for assessing causes of death in communities in which most deaths occur at home. The method is widely applied (Figure 1 and Table 1), and much effort has been put into developing and refining the approach for specific purposes.

This review summarizes methodological and conceptual developments, considering the VA process as a whole, from data capture to interpretation and data use. It also discusses current thinking and ways to characterize cause-specific mortality globally. Identification of the literature and the review process is illustrated in Figure 2. The remainder of this review is structured to follow the VA process, which includes data capture, cause-of-death categorization, interpretation, and questions of validity and application.

DATA CAPTURE

Many of the methodological and conceptual developments in VA have focused on data capture, where decisions about questionnaires, timing, interviewers, and respondents...
can have an important influence on interpretation and application of the results. However, as with most developments in VA, there has been little coordination between such activities, resulting in a confused picture overall and a failure to reach consensus.

**Questionnaires**

VA questionnaires should make use of all available information, including old medicine packets and hospital cards. Development of the tool should start by listing the causes that occur frequently, are likely to be derived by VA, are responsive to public health interventions, and are associated with recognizable symptoms (10, 12). Causes will vary by age group (13), and, although some authors have used frequencies of causes obtained from hospitals to illustrate the range of causes needed in the VA tool (12), such methods are subject to bias because the mortality profile of hospital deaths may be unlikely to represent the range and frequency of deaths in the community. Separate questionnaires for neonatal, child, maternal, and other adult deaths have been suggested (10, 14), and recent World Health Organization guidelines reiterate this recommendation (15). Nevertheless, a single VA instrument covering all deaths in all age groups can be used satisfactorily to prioritize public health problems, inform resource allocation, and target and evaluate interventions (16).

A number of different questionnaire formats exist. Open-ended formats are considered a more natural way for people to give information on the time sequence, severity, and duration of symptoms and may help to establish interviewer-respondent rapport (17, 18). Unlike highly structured questions, open-ended questions are considered more likely to include what is salient to the respondent (18). While some argue that minimal training is needed for the narrative probe to be added to VA tools (18), it is generally accepted that an open format requires more skilled interviewers and can increase variability between interviewers. Some authors report that purely open-format questionnaires require medically trained interviewers (14, 19–23). Furthermore, interviewers using open-ended questionnaires may introduce bias into data collection by recording histories that neatly fit into known or familiar disease descriptions or are based on the interviewer’s initial impression of causes of death (22). The potential for huge variation in length and complexity of accounts is a further concern, although similar techniques show that responses can be relatively short and lend themselves to content and sequence analysis, which may produce more useful and detailed information than that produced through closed questionnaires (18).

Closed questions provide information on symptoms that were absent or present, and they may be more likely than open histories to produce a complete inventory of signs and symptoms (17). A pure checklist format overcomes the need for specialist interviewers and assumes that more information will be recognized than recalled; however, strictly closed questions may not adequately capture all details of symptoms and may increase false-positives (14, 17, 19, 21, 22). Commonly, questionnaires make use of filter questions, but this design has the limitation of excluding a disease category if a false-negative is given (19).

The inherent limitations of each questionnaire design have generally led to a variety of tools being used. Some adopt only structured, closed questions (24), while others rely entirely on open-ended data capture (25). The advantages and disadvantages of open or structured questionnaires have been discussed but not formally assessed to our knowledge, leading to further confusion over appropriate methods and lack of standardization, which in turn limit comparability of results across settings (20, 26, 27). In general, questionnaires that combine open and closed questions are recommended, thus enabling chronologically ordered narratives to provide a fruitful adjunct to more structured methods (12, 14, 18, 28–30). With such a design, open-ended sections can be used for
internal checks of the validity of responses to closed ques-
tions (28). In practice, however, combined questionnaires
tend to be relatively long, and emphasis is often placed on
the open narrative, sometimes with closed questions com-
pleted less frequently than narrative sections (31).

Large-scale surveys can prohibit the use of long, detailed
questionnaires (20). Lulu and Berhane (21) call for further
simplification of the questionnaire and illustrate that tools
taking approximately 20 minutes to administer by lay field-
workers can derive adequate responses to all questions. Al-
though simplification may narrow the possibility of making
the very specific diagnoses claimed to be achieved by more
elaborate questionnaires, it may be necessary in some settings,
particularly when lay field-workers gather information on
deaths occurring long ago (21). Furthermore, brevity of the
interview and a good rapport with the respondent can contrib-
ute to high response rates and sustainability of VA projects
(22). Therefore, VA forms must be easy to use; it has been
recommended that, for cost and simplicity reasons, the forms
be limited to om or eth a n 2si de sof anA 4-s i zed s h e e to f
paper, with an emphasis on symptom checklists and only brief
narrative histories included (22). Advances in handheld com-
puter technology may simplify the data-capture process
further, and initial application in VA shows promise (32, 33).

Timing

The trauma of death may influence recollection of symp-
toms, and feelings of guilt during mourning may distort
accounts of events preceding death (34). As such, timing
of a VA interview may be an important determinant of the
overall outcome (14, 35, 36). However, decisions about tim-
ing are frequently based on convenience or convention rather
than evidence. The effects of recall may differ depending on
the context and the characteristics and demographics of the
deceased (14); people interviewed too early after a death
occurs may be reluctant to talk about it, whereas delaying
the interview for too long may result in problems recalling
symptoms or sequence (10, 27, 29, 37, 38).

Some studies suggest that recall does not affect reporting
of an event as tragic as the loss of life and that longer recall
is as reliable as short intervals (21, 26). With apparently no
significant improvement or depreciation of recall ability
with time, except for deaths during the neonatal period
(34, 39), as much time as possible should be allowed
between death and interview (35), and definition of a mini-
mum as well as maximum period from death to interview is
important (19, 23). Omission of very specific questions is
helpful in minimizing problems of recall, but long time
spans can affect the availability of respondents because of
outmigration or other reasons (21). Suggestions for optimal
delay range from 3 months (10) to 2 years (21, 40) after
death, with initial “condolence visits” being recommended
in some instances to prepare families for interviews (41),
although no standard procedures have been established or
suitably tested (31, 42).

Interviewers

It is important that interviewers be trained so they have at
least a basic understanding of the data they are collecting,
and such training would benefit from globally standardized guidelines (11, 14, 19, 26, 27, 35, 43). For example, increasing access to multimedia resources could offer opportunities to standardize understanding and recognition of common symptoms among interviewers in different settings. Nevertheless, training and tools that emphasize indigenous terminology and concepts of illnesses and their manifestations, rather than purely biomedical models, are crucial, and anthropological fieldwork may be an important starting point for understanding local concepts of mortality. A minimum level of mandatory training in counseling techniques for interviewers may also be appropriate because the psychological effects of VA on grieving relatives are unknown (44).

The educational background and gender of interviewers vary, possibly influencing the outcome of VA studies, although, to the best of our knowledge, little empirical investigation of this issue has been conducted (14, 23). Although it has been reported that well-trained lay people conducting culturally and linguistically sensitive VA interviews can obtain accurate information (28), others have advocated the use of medically trained interviewers or health professionals (11, 13, 23). The merits of different interviewers have never been formally studied, although the use of medically qualified interviewers is intuitively more costly and, with an increased tendency to interpret results during an interview, may affect repeatability (19). An “index of suspicion” on the part of the interviewer could overestimate diseases of interest, although there has been minimal evidence of this (38, 41). Ultimately, the preferred demographics of the interviewer will vary depending on the setting and the format of the questionnaire; in some instances, a mix of medically trained and lay interviewers may be most appropriate.

Use of many interviewers is operationally convenient but may produce unhelpful diversity in the results, with each interviewer gaining only limited experience in the VA procedure (23). It is generally accepted that interviewers should be from the same ethnic group as the respondents and, ideally, from the same village or region because words that designate diseases or symptoms may vary from place to place (10, 16). However, in many settings in which VA is used, it may be impossible to find interviewers with adequately high levels of education who can administer elaborate questionnaires (21), and local recruitment demands thorough consideration of ethical issues, such as maintaining confidentiality and ensuring adequate, formalized systems of informed consent and ethical approval (41, 44, 45).

Application of VA in multiethnic populations heightens linguistic and conceptual challenges, and free translations and multiple interviewers may be required to communicate effectively (39). Random reinterviewing for quality control purposes facilitates detection of methodological shortcomings and can motivate reliable fieldwork (22, 25), but repeating interviews is not always possible or appropriate given the sensitive nature of VAs (41). Thus, in all cases, the importance of interviewer training cannot be overemphasized.
Respondents

The process of identifying appropriate respondents for VA data capture is not formalized and may need to be standardized, taking cultural factors into account (14). Experience shows that close relatives rarely refuse to give information, with some surveys reporting 100% response rates (16, 43, 44). However, general reluctance to discuss death or moribund symptoms, and considerable resentment about completing lengthy questionnaires about deaths, particularly in view of insufficient medical services, have been reported in some settings (46, 47). In many settings, it is necessary to enroll respected community members to build awareness and acceptance of the unfamiliar nature of VA during the early stages of planning (22, 46).

Ranking respondent relationships to assess the effect of the respondent on the quality of VA data about maternal deaths, Hoj et al. (39) showed that male respondents provided more detailed information than women did and that the information was better if the respondent was present during the terminal illness (39). Inevitably, male respondents may be more common in patriarchal societies because women may be unwilling or unable to discuss VA issues with a stranger without a man’s consent. However, in a study of all-age and -sex mortality, Gajalakshmi et al. (29) concluded that wives appeared to be better responders than husbands were, perhaps because wives pay more attention to the health of males than vice versa (29). It is likely that the ideal respondent will vary depending on the specific purpose of the VA investigation; however, in practice, the most realistic approach may be to interview the household head but enable him or her to consult with others, especially women in the household (39). In many instances, it may not be appropriate or indeed possible to restrict VA interviews to a single respondent (19, 23, 36).

It has been shown that mothers are able to retrospectively report signs and symptoms of their child’s illness, but they may be unable to differentiate between degrees of severity (28). Deliberate denial of well-known signs and symptoms by the respondents, or the fact that certain conditions such as malnutrition are common and chronic in some settings, can mean that particular symptoms are more difficult to appreciate by the lay community and are therefore less easily measured (38). This latter point relates to one of the key underlying assumptions of VA, namely, that each disease category has a distinct pattern of symptoms that can be recognized, recalled, and reported accurately by lay respondents (30). There is an important distinction between recognition and recall abilities, and little empirical evidence exists regarding respondents’ ability to recognize symptoms, with most research focusing on recall ability (35, 48). Symptoms that are both poorly recognized and poorly recalled are likely to yield spurious results (35) unless some degree of weighting is applied to represent the reliability of reported indicators (49–51).

Recognition, recall, and reporting depends to a certain extent on cultural interpretation of symptoms (16, 52) as well as on social taboos and expectations (36). Phenomena based on Western medical concepts and definitions may play no part in a culture’s conception of illness and death, and it is sometimes unrealistic to expect lay respondents to distinguish between similar symptoms they consider equivalent (10, 13, 19, 39, 41). Local understanding is therefore essential if clinical and lay concepts and terms to describe symptoms are to coincide (52), and some authors have recommended the use of clinical photos to aid recall and reporting (30, 34).

ASSIGNING CAUSES

Cause categories and multiple causes of death

A persistent and perhaps limiting assumption is that, once captured, VA data should be interpreted and processed to mimic, as close as possible, clinical diagnoses and standard death certification. International standard death certificates call for a single underlying cause of death. However, ascertaining a single cause of death from various possible causes identifiable from the reported symptoms may be inappropriate using VA data, and the causal web found with many deaths does not fit the linear model of death certification. Evidence suggests that mortality is often due to the effects of multiple acute and chronic conditions. Selecting a single cause may be less appropriate than assigning multiple causes, and studies that focus on a single diagnosis to the exclusion of others remain subject to opinion and may lead to inaccurate estimates of the diseases associated with death and the potential benefits of intervention (11, 17–19, 23, 27, 28, 35, 38, 39, 53–55).

Assignment and interpretation of multiple causes requires further methodological thinking. Differential diagnoses, with deaths divided into proportions of contributing causes, is one feasible approach that accepts that it is not possible to derive a single diagnosis for those who have shared symptoms (35). This approach has been successfully applied in a number of settings, providing estimates of cause-specific mortality fractions (49, 51, 56). Criticisms that such approaches prohibit comparison with more clinical models are less important if VA specifically seeks to address epidemiologic needs at a population level, emphasizing reliable estimates of cause-specific mortality fractions for defined populations rather than certainty at the individual level.

An alternative approach is to use broad-category, consolidated causes, with multiple related causes assigned to a single category (54, 57). Such an approach is consistent with the philosophy that it is out of the question to analyze all possible causes of death through retrospective interviews and that, instead, VA should focus on broad care needs and cause categories of public health importance rather than traditional clinical and pathological approaches (11, 13, 49, 58). Reports of selective interventions reducing deaths to nontargeted diseases further support the use of consolidated cause categories (54). When used together, coding rules and consolidated analysis can identify and prioritize public health needs in settings in which deaths often occur as a result of multiple, interacting disease processes, any of which could be fatal and each of which needs different programmatic interventions (54).

Alternatively, hierarchical classifications may be used, as in the International Classification of Diseases (ICD) (59). A
hierarchy arranges causes of death in a particular order based primarily on the specificity of the causal definition or perceived certainty of each diagnosis, as well as the sensitivity of the causal definition and the presumed physiologic precedence of each cause’s contribution to a death (57). In principle, such standardized coding makes data internationally comparable; however, strict ICD principles are beyond the bounds of lay classification systems, requiring diagnostic testing and expertise that is often not feasible, and may be less appropriate than simpler cause groupings that support intervention and prevention activities (2, 10, 36, 45).

With a strong focus on addressing data gaps at the global level, the use of ICD enables recording, coding, and analysis of multiple causes while retaining comparability based on tabulation of a single underlying cause. The use of ICD principles may also reduce the proportion of deaths classified as indeterminate because, in the absence of sufficient evidence for a specific cause, they may enable assignment of symptom-based or ill-defined causes and syndromes within an anatomic organ system as a valid cause of death (60, 61). Nevertheless, coding rules and detailed training are important in determining the correct relation among underlying, intermediate, and immediate causes, and the ICD categories and coding rules used in VA need to be harmonized (14, 22, 55, 60–62). In practice, adapted hierarchical systems may be more appropriate, such as the Child Health Epidemiology Reference Group classification of causes of child deaths (63) based on the expected public health importance of individual causes of death, differing implications for intervention, and the ability to distinguish between the individual causes in low-resource settings (52, 53, 64). Use of hierarchies ultimately has limitations, however, because the cause distribution can vary widely depending on how the hierarchy is selected (12, 57).

**Deriving causes from VA data**

VA data need to be interpreted to derive probable causes of death. In some settings, the interviewer determines the cause of death at the interview stage; however, this method is considered to have poor validity, and diagnosis is recommended at a later stage (19, 23), generally using one of 4 methods of interpretation: physician review without algorithmic diagnostic criteria; physician review using algorithms; computer algorithms; and, more recently, probabilistic approaches.

**Physician review.** The most common interpretation method is local physician review of data without algorithms (14, 27, 57, 60, 65–67). Physician interpretation of questionnaires involves subjectivity and judgment. Therefore, it is customary that questionnaires be independently analyzed by at least 2 people; when there is disagreement, additional review is required and a consensus diagnosis is given. If no consensus can be reached, the death is recorded as “indeterminate” (10, 16, 30). Use of a generalist physician is considered better than a specialist if the purpose is to assess all-cause and all-age mortality (60), and open-ended, narrative data are considered particularly important for such expert judgment methods (17, 20, 43, 67, 68).

Extensive use of open-ended questions and the subjective nature of physician review has face validity because of its similarity to taking medical histories (69), but it means that specific techniques used for diagnosis can differ considerably between individuals over time and between settings (27, 67), and final diagnosis can sometimes appear arbitrary (70). Although sometimes disappointing (10, 49, 71), agreement of case-by-case and population patterns of causes of death between physicians reviewing VAs is generally reported as high or reasonable (23, 26, 30, 41, 43, 64, 68) and depends to some extent on physician training and specification of the number and detail of causes of death that physicians can derive.

These high levels of agreement are perhaps to be expected because they likely reflect consistency in physicians’ prior knowledge of local epidemiology (14, 41, 68). It has been shown that physician reviewers will not readily code diseases not expected in certain demographic groups (16) and often focus on the symptoms reported to be present, paying less attention to those that are absent (34). In some instances, physician reviewers show preference for highly specific diagnoses, imputing a specific etiology in the absence of adequate microbiologic data (56), and tend to make an unsubstantiated selection of a single cause even if multiple causes are indicated (14, 51). Such nuanced interpretation introduces bias, particularly for less obvious causes of death for which decisions must sometimes be made between equally likely causes (56). If training results in bias toward a particular cause-death assignment, the skew is likely to be repeated by subsequent coders because the training and supporting materials provided within a VA system will be similar (68). As such, expensive steps of multiple review and consensus may not be justified, at least for some diagnoses (36, 68).

Disagreement between physicians most commonly concerns a binary outcome, such as whether or not a woman had signs of eclampsia (26). Such discrepancies arise during individual-level cause-of-death ascertainment, but they appear to have a minimal impact on overall cause-specific mortality patterns at the population level, where misclassifications of causes appear to be bidirectional (64, 68, 71). Nevertheless, comparison of physician-derived results between settings or different points in time is hindered by concerns over repeatability (17, 23, 26, 69). Furthermore, the use of physician reviewers is expensive, requiring from 5 minutes (56) to half an hour (12, 66) of physician time per VA case and diverting medical practitioners from clinical roles in settings in which physicians are scarce, which prohibits the use of this method in large-scale studies (17, 68, 71).

**Algorithms.** Long-standing concerns about interobserver agreement and lack of standardization of physician review methods have been addressed with efforts culminating in the development of various algorithmic approaches based on the concept of distilling the process of physician review into standardized rules. Algorithmic approaches are better for comparability between time and place and essentially use combinations of specific symptoms to diagnose specific causes of death (20, 27, 39). Algorithms are closely linked to the design of questionnaires and field procedures, and their use is limited by the availability of necessary indicators.
in a setting. The discriminatory power of indicators also depends on their prevalence in the study setting (20).

Algorithms can be developed from textbook descriptions, existing clinical algorithms, local experience, or a combination of all 3, and the process of deriving algorithms varies and is not always specified in studies that use them (19). Use of local clinical experience in creating “expert algorithms” can be considered to represent a consensus of physician reviewers (69) but hinders their use in other settings. Textbooks are sometimes used to derive algorithms but may not be appropriate in some settings because of differences in cultural perceptions of symptoms of disease (19). Alternatively, data-derived algorithms can be developed whereby cause-specific symptoms may be found by using standard statistical techniques such as cross-tabulations, logistic regression, decision trees, and Bayesian classification, which identify discriminant functions of different indicators (20, 69). Artificial neural networks that apply nonlinear statistics to pattern recognition have also been proposed, but they are more complex and have questionable generalizability (67).

A major limitation of data-derived methods is dependency on the availability of true or medically confirmed causes. A proportion of known mortality data is used to develop the algorithm and obtain the highest possible levels of sensitivity and specificity (the “train” data set), and the remaining data are used to assess the performance of the algorithm (the “test” data set) (20, 24, 66, 67, 71). Algorithms have been reported to exclude certain cause-of-death possibilities when health facility records do not show these conditions to be major causes of hospital admissions or death in the hospital (21), but the assumption that hospital deaths are representative of community-based deaths for which VA will be used is flawed. Furthermore, data-derived algorithms might discriminate well between causes of death only in the population from which they are derived and therefore may have poor repeatability, which is a particular problem for causes of death for which numbers are small (66).

A major difference between expert algorithms and data-derived methods is that the latter use indicators that best discriminate between causes and not necessarily the symptoms that physicians deem to be essential, confirmatory, or supportive in clinical diagnosis, irrespective of whether the sign discriminates well between different diseases (20, 69). This method of designing diagnostic techniques may be more appropriate; as previously described, some symptoms, which medical knowledge would suggest are highly specific, have context- and culture-specific interpretations, which diminishes their diagnostic importance in some settings (72). Diagnostic criteria of algorithms may seem insufficient by clinical standards, yet simpler algorithms are preferable and often perform better (28, 73), perhaps because the diagnostic criteria are applied in situations in which fatal outcomes have already occurred, imparting the criteria with a significance different from that of clinical interpretation (11).

Separate algorithms need to be developed for each cause, and, whereas they avoid the subjectivity of physician review, rigid and highly specific diagnostic criteria are likely to lead to indeterminate cases in which physician review may have been able to reach a diagnosis (39, 57, 74). Using different algorithms for each cause and thereby separately checking for multiple causes of death is possible and, given adequately high specificity of each algorithm, ensures that deaths are not assigned too many causes (20, 66, 71). Hierarchical algorithms with clearly defined rules for distinguishing between underlying and immediate causes can be applied in such cases (12, 39, 48), but the detail needed to do so may be unrealistic and, because of the finite list of indicators and causes, makes algorithmic approaches less amenable to measuring all-age and all-cause mortality in different settings in which a variety of causes might be expected (20, 23). The diversity of VA questionnaires also limits application of standardized algorithms across different settings (48, 71), and context-specific adaptations may be necessary (11). It is unlikely that data-derived algorithms are good at assigning causes for some deaths that tend to have nonspecific symptoms, such as cardiovascular disease, pneumonia, or malaria, (27, 71), and the difficulty of considering parallel possibilities similar to clinical differential diagnoses is a further limitation of algorithmic approaches (50).

In general, algorithms do not use data generated from open-ended questions (20). Exclusion of information considered valuable to physicians makes algorithms simpler to use and less subjective (71) but has somewhat limited their acceptance because of the prevailing opinion that open-ended questions are somehow more important than closed ones (14, 43, 66). Nevertheless, there is no evidence that adding symptoms coded by using the verbatim sections of questionnaires improves the performance of algorithmic approaches, probably because respondents rarely provide additional information in these open-ended sections but instead subjectively emphasize symptoms they consider most important (48). Despite this limitation, where physicians are guided by algorithms but are free to assign alternative causes of death, relying exclusively on structured, closed questions precludes unanticipated data that might inform the final diagnosis, and verbatim sections apparently provide information that can lead physicians to assign causes not included in the algorithms (36, 48).

**Probabilistic approaches.** Whereas algorithmic approaches to VA interpretation assess the presence or absence of single causes based on absolute positive or negative responses to symptom-related questions, automated methods that apply probabilistic reasoning adjust the probability of a range of multiple possible outcomes simultaneously, dependent on population-level and cause-specific probabilities of specific symptoms being reported given a specific cause of death. Thus, algorithmic methods result in a binary outcome (yes or no) for a single cause, whereas probabilistic methods quantify the probability of a range of multiple causes.

Application of Bayes’ theorem for such probabilistic interpretation of VA data has been developed and evaluated in various settings (49, 50, 56). Establishing prior probabilities needed in a Bayesian model was, for some time, considered too complex (8, 20). Use of facility-based data would be subject to the inherent biases of applying hospital-based approaches to quantify community-based symptom patterns, and establishing discriminatory power would be particularly challenging for deaths that share common symptom complexes that, in a clinical setting, would enter
a differential diagnosis process (35). However, recent work has shown that a high degree of precision in Bayesian probabilities is not required to achieve a workable VA model that estimates population-level cause-specific mortality patterns, and expert consensus on underlying probabilities has been used successfully (50, 51).

Simultaneously adjusting the probability of each of a finite list of causes according to affirmative answers to specified symptoms, the Bayesian “InterVA” approach calculates the likelihood of each cause and displays as many as 3 of the most probable causes, along with their associated likelihoods. A certainty indicator is also calculated, which may help with interpretation of results and fits with opinions that identifying possible causes and degrees of certainty of derived causes may be more useful than definitive answers in relation to VA data (11). By assigning multiple causes with associated probabilities, the InterVA approach may provide a solution to the long-standing VA dilemma of quantifying a form of differential diagnosis process for deaths from common symptom complexes while not relying on algorithms (57).

The InterVA method has been shown to produce VA-derived, cause-specific mortality fractions comparable to physician review, with the advantage of being a completely reliable and standardized approach to interpretation, allowing certain adjustments to reflect local disease prevalence (49, 51). The method overcomes concerns regarding physician reliability and is amenable to data from any questionnaire design, including open-ended sections, although information reported exclusively in these sections appears to add little to the model’s performance (49).

The efficiency and usability of InterVA is attractive, and its standardized interpretation method has been applied in intervention and impact evaluation research in different settings (32, 75–77). Although statistical modeling of this sort may not reflect the subjective subtleties of physician review (8), it is advantageous in terms of efficiency, consistency, and standardization. Nevertheless, the method is considered of limited use at the individual level, and the lack of a “gold standard” against which to validate diagnoses has somewhat restricted its application (14).

An alternative probabilistic approach developed by King and Lu (78) directly estimates cause-specific mortality fractions without individual cause-of-death attribution, and it resolves the problem of generalizing VA analysis to the population based on test properties quantified in facility-based validations. Data on symptoms reported by caregivers, along with the cause of death, are collected from health facilities, and the cause-of-death distribution is estimated in the population in which only symptom data are available. This method is more complex than InterVA and, with theoretical results and empirical analyses using data from China and Tanzania, provides a more theoretically sound approach to setting probability levels. However, it is limited in that it depends on the availability of high-quality, facility-based or valid mortality data, and such data simply do not exist in the majority of settings in which VA is needed. King and Lu’s work provides a refreshing focus on estimating aggregate cause-specific mortality proportions, describing them as the quantities of primary interest in VA studies.

Combining King and Lu’s (78) approach with InterVA, Murray et al. (72) propose, and have attempted to validate, a new “symptom pattern” method. First, this method requires a data set for which the true cause of death is known but which need not be representative of the population of interest (such as death occurring in hospitals). From this, the probability of reporting specific symptoms given a specific cause can be quantified. These symptom properties then allow population- and individual-level cause patterns to be determined from VA data from a second data set from the population of interest (72). The symptom pattern method addresses many issues and, based on application in China, shows promise. Ultimately, however, this method, as with King and Lu’s, requires large databases of symptom-level sensitivities and high-quality, facility-based cause-of-death data—a highly context-dependent prerequisite that cannot readily be met by the majority of settings in Africa and elsewhere that need to use VA methods. Both approaches address important information gaps at the global level and represent a significant step forward in VA methodology and conceptual thinking. However, gaps remain, and it is becoming increasingly clear that gaps in mortality data that are most important to fill cannot all be met by a single methodology (2).

VALIDATIONS

Just as the VA process has many components, its validity is influenced by the real cause-of-death pattern and characteristics of the deceased, and by several other factors relating to the design and content of the questionnaire, field procedures, and interpretation of data (19, 26, 48). Most validations have compared VA diagnoses for a selected group of causes with clinical diagnoses, hospital records, or death certificates (13, 28). Such comparisons may be useful, but the selective nature of the study population and imperfect physician diagnoses prohibit a complete assessment of validity (28).

The use of death certificate information is problematic; even when they are available, certifiers often report nonspecific diagnoses and may be unreliable (13, 28, 29, 39, 43, 55). Facility-based validations are also inadequate, with concerns regarding the quality and quantity of information available in medical records, particularly in rural areas, and hence the acceptability of clinical diagnoses, which have been shown to differ from actual postmortem diagnoses (20, 28, 30, 31, 38, 79, 80). The highly selective nature of health service use in most low-income settings is a further limitation of using facility-based data to validate a community-based tool, and results are likely to represent a best-case scenario (37, 74). Previous contact with health services or in-hospital deaths may affect the ability of respondents to recall and report symptoms and cause of death, while the natural history of deaths, performance of VA, and diagnostic biases may be different in community compared with hospital samples; indeed, if it were not, then there would be little need for VA because extrapolation from hospital data would suffice (13, 16, 27, 30, 36, 55).
“Validation” studies are often limited to a small number of diseases and inadequate sample sizes, and results tend to be considered in terms of sensitivity and specificity—measurements that depend not only on questions asked and diagnostic methods used but also on the prevalence of diseases with overlapping symptoms and local awareness or}

Table 2. Mean, Minimum, and Maximum Sensitivities and Specificities for Verbal Autopsy for Selected Causes of Death Available in the Published Literature

<table>
<thead>
<tr>
<th>Cause Category</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean %</td>
<td>Range</td>
</tr>
<tr>
<td>Abortion</td>
<td>60</td>
<td>56–64</td>
</tr>
<tr>
<td>Acute abdominal disease</td>
<td>58</td>
<td>37–92</td>
</tr>
<tr>
<td>Acute febrile illness</td>
<td>46</td>
<td>0–74</td>
</tr>
<tr>
<td>Acute respiratory infection</td>
<td>54</td>
<td>28–86</td>
</tr>
<tr>
<td>Anemia</td>
<td>26</td>
<td>17–33</td>
</tr>
<tr>
<td>Cardiovascular diseases</td>
<td>47</td>
<td>25–87</td>
</tr>
<tr>
<td>Cerebral malaria</td>
<td>66</td>
<td>39–94</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>59.7</td>
<td></td>
</tr>
<tr>
<td>Congenital abnormality</td>
<td>60</td>
<td>40–76</td>
</tr>
<tr>
<td>Diabetes</td>
<td>58</td>
<td>43–75</td>
</tr>
<tr>
<td>Diarrheal diseases</td>
<td>60</td>
<td>0–100</td>
</tr>
<tr>
<td>External causes</td>
<td>88</td>
<td>35.3–100</td>
</tr>
<tr>
<td>HIV/AIDS</td>
<td>53</td>
<td>31–76</td>
</tr>
<tr>
<td>Liver diseases</td>
<td>29</td>
<td>0–71.4</td>
</tr>
<tr>
<td>Low birth weight</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Malaria</td>
<td>46</td>
<td>0–94</td>
</tr>
<tr>
<td>Malnutrition</td>
<td>64</td>
<td>0–100</td>
</tr>
<tr>
<td>Measles</td>
<td>82</td>
<td>54–98</td>
</tr>
<tr>
<td>Meningitis</td>
<td>52</td>
<td>22–93</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>52</td>
<td>6–96.1</td>
</tr>
<tr>
<td>Noncommunicable diseases</td>
<td>66</td>
<td>52–75</td>
</tr>
<tr>
<td>Other digestive disorders</td>
<td>51.8</td>
<td></td>
</tr>
<tr>
<td>Other infections</td>
<td>71</td>
<td>39–86</td>
</tr>
<tr>
<td>Perinatal asphyxia</td>
<td>60</td>
<td>43–87</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>47</td>
<td>9–80</td>
</tr>
<tr>
<td>Pregnancy-related causes resulting in death of the mother or child</td>
<td>60</td>
<td>11–100</td>
</tr>
<tr>
<td>Prematurity</td>
<td>65</td>
<td>48–87</td>
</tr>
<tr>
<td>Rabies</td>
<td>94</td>
<td>83–100</td>
</tr>
<tr>
<td>Renal diseases</td>
<td>28</td>
<td>0–60</td>
</tr>
<tr>
<td>Sepsis</td>
<td>44</td>
<td>6–75</td>
</tr>
<tr>
<td>Stillbirths</td>
<td>71</td>
<td>61–80</td>
</tr>
<tr>
<td>Tetanus</td>
<td>61</td>
<td>17–100</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>54</td>
<td>18–92</td>
</tr>
<tr>
<td>Tuberculosis/AIDS</td>
<td>58</td>
<td>0–91</td>
</tr>
</tbody>
</table>

Abbreviations: AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.
that accuracy in one aspect—say, cause-specific mortality distributions—does not necessarily correspond with accuracy estimated by other methods, such as kappa, sensitivity, or specificity (12, 16, 22, 27, 48, 60, 61, 66, 67, 81).

True validation of VA against hospital-based data ultimately fails (46, 56, 60), and an absolute gold-standard assessment of cause of death that can be applied to deaths at home is unlikely to exist (66). Nevertheless, VA diagnoses correlated with hospital diagnoses continue to be considered useful in illuminating some of the limitations of VA, as long as they are interpreted carefully and it is recognized that such comparisons can, at best, show that VA-derived causes are as good as those from hospital physicians for a selected, and potentially biased, proportion of all deaths (12, 74)—an assessment that must not be confused with validity. Such comparison can also be useful in identifying obvious miscategorization errors, but, even with sophisticated statistical procedures (82), it is unlikely that the effect of misclassification can be adjusted for in the absence of adequate reference data (14, 65, 81, 83). Instead of discussing absolute validity, therefore, assessments of VA should focus on plausibility and usefulness in filling important gaps in mortality data at the community level to address specific public health needs, perhaps with a more qualitative assessment of the robustness of results and “cultural validation” for appropriateness (22, 46, 60).

### Table 3. Empirical Research Needs in Verbal Autopsy

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Unanswered Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaires</td>
<td>What are the benefits and limitations of different questionnaire designs on overall completeness, quality, and utility of VA data as well as on operational costs and effects on respondents? Is there a role for direct data capture methods using handheld computers in VA?</td>
</tr>
<tr>
<td>Timing</td>
<td>How does timing affect the completeness and quality of VA data? Are recall and reporting abilities affected by time and in the same way for all deaths? How are ethical issues of population research incorporated into the design and implementation of VA methods?</td>
</tr>
<tr>
<td>Interviewer effects</td>
<td>How do characteristics of the interviewers affect the completeness and quality of VA data? How can interviewer training be harmonized to facilitate standardized VA data capture methods globally?</td>
</tr>
<tr>
<td>Respondents</td>
<td>What are the characteristics of “ideal” respondents in terms of accurate recall and reporting ability? How can local understanding of health and illness be incorporated into VA and used to inform cause-of-death ascertainment? What emotional and psychological effects does VA have on grieving respondents?</td>
</tr>
<tr>
<td>Cause categorization</td>
<td>How should multiple possible causes of death be handled? How can coding rules, consolidation of causes, and hierarchical classification systems be adapted to VA and standardized globally?</td>
</tr>
<tr>
<td>Assigning causes, evaluations, and applications</td>
<td>How well do different VA interpretation methods perform in relation to explicit data needs? How can methods be further developed and refined with respect to these data needs? In the absence of “gold standards,” what criteria should be used to assess VA methods? How can VA be adapted beyond the traditional medical paradigm of cause of death to incorporate broader determinants of mortality?</td>
</tr>
</tbody>
</table>

Abbreviation: VA, verbal autopsy.

Each of these aspects is interrelated, and methodological research and conceptual developments should not consider any one aspect in isolation.

An alternative measure of validity has been kappa statistic assessment of agreement between differing methods of interpretation, thereby avoiding the flawed assumption of sensitivity, specificity, and predictive value calculations that one specific diagnosis is an absolute gold standard (50). Comparison of cause-specific mortality fractions is a further useful means of assessing the performance of VA methods in terms of plausibility and adequacy for purpose at the population level, with estimates within 20%–25% of one another deemed acceptable (49, 54, 56, 60, 61, 66, 71, 81). Likelihood ratios of false-positives to false-negatives can also be used to indicate the degree of misclassification resulting from different VA methods, although these measures once again rely on the availability of absolute gold-standard diagnoses (70). Recently, studies have combined the measures of validity described above, acknowledging recognition (12, 16, 17, 19, 20, 28, 36, 38, 74). A summary of published validation results in terms of sensitivity and specificity is presented in Table 2. Inclusion of a group of syndromes not likely to cause diagnostic confusion may be unduly optimistic because specificity will be greatly increased and predictive values may be more relevant (30). Although various arbitrary levels have been suggested, there is no universally recognized cutoff above which the levels of sensitivity, specificity, and predictive values are deemed acceptable, and the minimum levels depend on the intended use of the results (12, 55, 60, 66).

An alternative measure of validity has been kappa statistic assessment of agreement between differing methods of interpretation, thereby avoiding the flawed assumption of sensitivity, specificity, and predictive value calculations that one specific diagnosis is an absolute gold standard (50). Comparison of cause-specific mortality fractions is a further useful means of assessing the performance of VA methods in terms of plausibility and adequacy for purpose at the population level, with estimates within 20%–25% of one another deemed acceptable (49, 54, 56, 60, 61, 66, 71, 81). Likelihood ratios of false-positives to false-negatives can also be used to indicate the degree of misclassification resulting from different VA methods, although these measures once again rely on the availability of absolute gold-standard diagnoses (70). Recently, studies have combined the measures of validity described above, acknowledging
APPLICATION

There are different levels at which cause-specific mortality patterns are needed, and for which VA methods are applied. At the local level, health managers need timely and accurate mortality data that can be derived by using accessible and efficient methods to effectively plan health services based on local patterns of disease (73, 84). Epidemiologists, researchers, and assessors of specific health interventions need consistent assessments of cause-specific mortality to determine trends and evaluate interventions across regions and time (17, 30, 85). National and global authorities require standardized cause-specific mortality data from a wide range of settings to monitor and compare disease burdens (2, 58). It is perhaps unrealistic to expect any one approach to satisfy data needs at all levels (16, 36, 55, 64, 66, 71, 85), and, first and foremost, VA is a public health tool (19, 21, 36, 50, 56, 70).

In VA literature, a confused picture persists, with strong individual-level focus on methods and population-level focus on application of results (22, 37). Ambiguity regarding the ultimate purpose of VA methods can result in a narrow assumption that the approach is a direct surrogate for Western-style cause-of-death determination, which undermines its potential as an epidemiologic tool that can be adapted to any point along the chain of economic, social, operational, biomedical, and physical events leading to death, including failures in health and social services (2, 18, 39, 57, 86, 87). VA can be designed to address specific public health or mortality questions in a way that Western, medical-based models cannot, and explicit targeting of a specific point in the chain of events leading to death is useful in terms of data collection, analysis, and application. Ultimately, use of VA should be based on locale-specific studies and an appreciation of the limitations inherent in the technique, and, in certain situations, alternative approaches to measuring cause-specific mortality should be sought (17, 30, 35).

CONCLUSIONS

There remains relatively little in the way of empirical evidence supporting what have become routine VA procedures. All aspects of VA should be based on empirical evidence, and no single aspect of the process should be considered in isolation (Table 3). Research and development of the data-capture aspects of VA has received much attention (15) and has been important, but it will not prevent questionable variation in cause-proportionate mortality patterns if subsequent steps in the VA process remain nonstandardized (62). Advances in methods of interpreting VA data and categorizing causes are likely to benefit from a greater population-level focus and broader conceptualization of cause of death in public health terms as well as in traditional medical models. Maintaining a structured, quantitative, and purely biomedical framework of the most common causes of death risks excluding important health constructs that could allow monitoring and intervention on multiple levels (biomedical, social, and cultural) (88). VA cannot be reduced to a single one-size-fits-all tool. It must be contextu-

alized; given the lack of gold standards, methodological developments should not be considered in terms of absolute validity but rather in terms of consistency, comparability, and adequacy for the purpose (40, 89). Ultimately, clarity and consistency about what VA methods are trying to achieve is crucial and this demands collaboration and harmonized efforts.

ACKNOWLEDGMENTS

Author affiliation: Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden.

This work was undertaken within the Umeå Centre for Global Health Research, with support from FAS, the Swedish Council for Working Life and Social Research (grant 2006-1512).

Conflict of interest: none declared.

REFERENCES


203. Fikree FF, Azam SI, Berendes HW. Time to focus child survival programmes on the newborn: assessment of levels


