Prospective Study of One Million Deaths in India:

Rationale, design and validation results

RGI-CGHR Prospective Study Collaborators

Final Version, September 30, 2005

Correspondence to:

Dr. Prabhat Jha, Director and Canada Research Chair of Health and Development
Centre for Global Health Research, Public Health Sciences, St Michael's Hospital, University of Toronto
70 Richmond Street East, 3rd Floor, Toronto, Canada, M5C 1N8
Phone (416) 864-6042
Fax (416) 864-5256
Email prabhat.jha@utoronto.ca

Word count : Main text 6,133, Abstract:393
Figures: 4
Tables: 3
References: 54
Supplemental text : 2,959 words, 15 references

Running title: Million Death Study in India

Abstract

Background: Over 75% of deaths in India occur in the home; more than half of these do not have a certified cause. India and other developing countries urgently need reliable quantification of the causes of death. They also need better epidemiological evidence about the relevance of physical (e.g. blood pressure and obesity), behavioral (e.g. smoking, alcohol, HIV-1 risk taking and immunization history), and biological (e.g. blood lipids and gene polymorphisms) measurements to the development of disease in individuals or disease rates in populations. We report here on the rationale, design and implementation of the world's largest prospective study of the causes and correlates of mortality.

Methods: We will monitor nearly 14 million people in 2.4 million nationally-representative Indian households (6.3 million people in 1.1 million households in the 1998-2003 sample frame and 7.6 million people in 1.3 million households in the 2004-2014 sample frame) for vital status, and if dead, the causes of death through a well-validated verbal autopsy (VA) instrument. About 300,000 deaths from 1998-2003 and some 700,000 deaths from 2004-2014 are expected; of these about 850,000 will be coded by two physicians to provide causes of death by gender, age, socioeconomic status and geographical region. Pilot studies will evaluate the addition of physical and biological measurements, specifically dried blood spots.

Results: Preliminary results from over 30,000 deaths suggest that VA can ascertain the leading causes of death, reduce the misclassification of causes and derive the probable underlying cause of death when it has not been reported. VA yields broad classification of the underlying causes in about 90% of deaths before age 70. In old age, however, the proportion of classifiable deaths is substantially lower. By tracking underlying demographic "denominators" the study permits quantification of absolute mortality rates. "Household case-control", "proportional mortality" and "nested" case-control methods permit quantification of risk factors.

Significance: This study will reliably document not only the underlying cause of child and adult deaths, but key risk factors (behavioral, physical, environmental and eventually, genetic). It offers a globally-replicable model for reliably estimating cause-specific mortality using VA, and strengthens India's flagship mortality monitoring system. Despite the misclassification that is still expected, the new cause of death data will be a quantum better than that available previously.

Introduction

About 46 million of the estimated 60 million deaths worldwide occur in developing countries (1). However, there is a dearth of reliable and accurate information on the causes and distribution of mortality in these countries. India has about nine million deaths a year, about one in six of all deaths worldwide. Over three quarters of deaths in India occur in the home; more than half of these do not have a certified cause.

To meet these modern challenges of mortality measurement, the world's largest prospective study of the causes and correlates of mortality in India is being undertaken by the Registrar General of India (RGI)'s Sample Registration System (SRS). The study, called the RGI Million Death Study in India, is implemented in close collaboration with the Centre for Global Health Research at the University of Toronto, leading Indian and overseas academic institutions and the Indian Council of Medical Research. The study has several innovations that are relevant to other developing countries considering the measurement of mortality, and to recent calls for improved health statistics (2-5). It uses a well-validated household instrument to ascertain causes of death and dual recording methods to improve reliability and consistency. It is national in scale, representative of the population and, by recording underlying demographics, able to quantify absolute mortality rates. Prospective and retrospective study designs such as nested-case control, household and proportional mortality methods permit quantification of correlates of mortality.

In this report, we present the rationale, design, results of validation studies to date, key statistical issues, and expansion to biologic measurement for the Million Death Study in India. We discuss the challenges in implementing modern mortality measurement, and the implications for global health.

Rationale: Why measure mortality?

Historically reliable, representative, routine, low-cost and long-term mortality measurements are the key to monitoring trends in health conditions of the population, detecting new epidemics (such as HIV/AIDS), spurring research into avoidable causes of death, evaluating the success of control programs and improving accountability for expenditures on disease control (6,7). Routinely collected data have helped to spur further research and public health action and contributed to the enormous increases in life expectancy in the 20th century (8).

Public health in industrialized countries was transformed when vital statistics on age, sex and socioeconomic distribution of births and deaths became available in the late 19th and 20th centuries. Vital statistics have demonstrated major trends in fertility, child survival and mortality. They have shown good news, such as the large declines in under-five mortality and tuberculosis mortality during the 20th century. They have also raised alarm; in the mid 1940s, a dramatic increase in lung cancer deaths in British and American men after World War II led to much research on smoking (9). In the early 1980s, routine mortality data from San Francisco revealed an exceptional increase in immune-related deaths among young men and signaled the start of the American HIV-1 epidemic (10).

Vital statistics need to keep up with modern patterns of disease. India and other countries have seen consistent decreases in child mortality (under-five mortality has fallen by about 2% per year since 1971 in India; 11). Adult deaths in middle age (35-69 years) attributable partially to the effects of smoking, sedentary lifestyles and higher saturated fat intake have been on the rise. More recently, deaths among young adults (15-34 years) have risen from HIV/AIDS. Reliable information on the diseases of adults and their causes are a large gap in global knowledge.

Recorded deaths from most communicable diseases or injuries generally corresponded to their causes (e.g. malaria deaths are caused by plasmodium parasite), but the more “chronic” communicable diseases (such as tuberculosis) and most non-communicable diseases can have multiple causes. For example, a myocardial infarction could be caused by smoking, elevated blood pressure, high lipids or other factors. The age and sex-specific importance of established risk factors, or combinations of risk factors, has only recently been reliably documented through appropriately large studies in Western populations, and with surprising results. For example, the association of blood pressure and vascular disease is twice as steep as previously believed (12) if blood pressure is measured reliably and the effects of “regression-dilution” bias (13) are properly considered. There are few epidemiological studies that document the age, sex, and region-specific hazards of blood pressure, blood lipids and smoking in developing populations (14), where most of the world’s vascular deaths occur (1) .

Anecdotal evidence suggests that in India each disease that is common in one part of the country is relatively uncommon elsewhere, for reasons that are not understood. This means that there are important avoidable causes that still await discovery. Much more remains to be

discovered about the novel genomic, proteomic and other biochemical correlates of respiratory, intestinal or other infections in general, and of the avoidable causes of chronic diseases such as cancer, heart attack, stroke and lung disease (15;16) that currently account for most of the adult mortality in India. Even for infections such as HIV-1 and tuberculosis, there may well be genetic causes (such as polymorphisms in genes involved in innate (17;18) or adaptive (19) immune recognition) or environmental causes (such as other infections (20;21)) other than the relevant pathogen that make particular infections, or progression from infection to disease, more probable.

Alternative designs for measuring mortality

The ideal mortality measurement system has several characteristics (7). It is routine, reproducible, long-term, low-cost and sustainable. It is reliable and representative of the population (implying that it avoids major selection biases in enrolment). It captures not only the death, but also reliably the cause based on the International Classification System of Diseases (ICD-10; 22). It includes not only events or "numerators" but also underlying population demographics ("denominators").

Three systems measure mortality in India (www.censusindia.net). The first, the Civil Registration System, is currently unreliable due to gross under-registration. While some areas have very good vital registration (Mumbai provides death registration as far back as 1848 (23)), overall, only 2.8 million of the estimated 9.5 million annual deaths were registered in India in 1995. Among registered deaths, cause-of-death data are available for about one in three deaths, but this often merely subdivides deaths as due to accident, violence or disease, without further details. Civil registration is the accepted mortality measurement system in Western countries where coverage nears 100%. However, access to medical care is far less common in India, and most deaths occur at home rather than in hospitals. Eventually, civil registration will increase, as has happened recently in China. But this may take decades; the United States took the better part of a century to increase death certification, and some states did not have complete coverage until the 1970s (24).

The second system is the Medically Certified Causes of Death. However, this covers only about 0.4 million deaths, and is largely confined to selected urban settings that are not representative of the general population. Problems with inconsistent physician attribution of causes of death, especially for senility and ill-defined causes have been noted (6,7).

The third is the Sample Registration System (SRS; described more below). Only the SRS is representative of both urban and rural settings of India, covering some 6,700 to 7,600 units randomly selected from the preceding census. The SRS is much smaller, though, covering only 0.05 million deaths. Thus, its chief draw back is that it cannot yet provide district-level data for local decision making, and lacks sufficient power to generate yearly rates for less common causes, such as causes for maternal deaths. Until recently, the SRS did not adequately capture information on causes of death. However, we have addressed this gap by developing and implementing verbal autopsy (VA) - an innovative method to estimate cause-specific mortality (see description below).

Finally, it is worth noting that econometric models of cause-specific mortality (1) are no replacement for direct measurement. Indirect estimates are only as good as their underlying data. The econometric models have not been well tested in the presence of HIV/AIDS growth. In India the global burden of disease varies considerably depending on the assumptions used. For example, the 1994 global burden of disease estimated 0.78 million cancer deaths in 1990, but registry data suggested a much lower figure of 0.43 million deaths (25). The 1996 version of the global burden of disease projected 0.95 million deaths from tuberculosis deaths in 2000; the 1999 version estimated the number of deaths in 1998 at 0.42 million (26).

Study objectives

In light of the above, we have the following objectives:

- Reliably document cause-specific mortality from 2001 to 2003 (three years; ~150,000 deaths) within the SRS to establish regional, gender, and age-specific variation and patterns of mortality.
- Document causes of death with routine use of VA in the new SRS sampling frame from 2004 to 2014 (ten years; ~700,000 deaths expected).
- Improve our understanding of selected risk factors, most notably tobacco use, alcohol use, indoor air pollution, fertility preferences for male children and its effect on female survival, immunization and migration through linking of mortality outcomes with exposure status using retrospective and prospective methods.
- Expand the new SRS to obtain reliable epidemiological evidence about the relevance of physical (such as blood pressure and peak flow), behavioral (such as migration and HIV-1

risk), and biological (such as blood lipids and gene polymorphisms) measurements to the development of disease in individuals or disease rates in populations.

Methods

Study setting

This study is conducted within the SRS, a large, routine demographic survey and the primary system for the collection of Indian fertility and mortality data since 1971 (27). There are two SRS sample frames. The first SRS sample frame covers 6.3 million people (including 2.9 million adults aged 25 years or above) in all 28 states and seven union territories of India. An average of 150 households are drawn from each of 6,671 sample units (4,436 rural and 2,235 urban), which in turn are selected using 1991 Census data. The new SRS sample frame covers about 7.6 million people (including 3.5 million adults aged 25 years or above) in all 28 states and seven union territories of India. Households are drawn from 7,597 sample units (4,433 rural and 3,164 urban) selected from the 2001 Census.

SRS sample units are randomly selected to be representative of the population at the state level. The sample design is a uni-stage stratified simple random sample without replacement. The sample size for the first and new SRS sample frames are based on total fertility rates and infant mortality rates, respectively. Within the SRS, selected households are continuously monitored for vital events by two independent surveyors. The first is a part-time enumerator (commonly a local school teacher familiar with the area/village) who visits the home every month. The second is a full-time (non-medical) RGI surveyor who visits the home every six months. Each independently records the births and deaths in the household for a six-month period. A third staff member does a reconciliation of the two reports, arriving at a final list of births and deaths for each household, which completes each half yearly survey. The RGI surveyors each cover about 150 households with a total average population of 900 (ranging from 700 to 1,500), and report about 50 deaths every six months.

Plan of investigation

Box 1 and Figure 1 provide an overview of the study methods. In brief, the method involves 800 trained (non-medical) RGI surveyors implementing VA reports among enrolled populations every six months. A random 10% of the VA field work is repeated by an independent audit team. After data entry, field reports are sent electronically to two independent and trained

physicians who assign cause of death based on the International Classification of Disease, 1990 (ICD-10) using a web-based system. The two physicians have to agree on the underlying cause of death and if they do not, such records undergo reconciliation and third-physician adjudication.

[Insert Figure 1] / [Insert Box 1]

The supplement information (Annex 1) provides the details of the field collection methods, re-sampling, physician coding, data management and research ethics. The full protocol, field instruments, training manuals, slide presentation of validation results and procedures are available at www.cghr.org/project.htm.

Estimated sample size & distribution

The primary outcomes of interest are all-cause mortality and cause-specific mortality. We expect about 150,000 cause-specific deaths determined through VA and about twice as many deaths (300,000) to study all-cause mortality in the 1998-2003 sample frame. The expected age distribution of cause-specific deaths captured with VA, based on the age-specific SRS death rates in 2001, is shown in Figure 2. Using indirect WHO estimates on causes of death in India (1), we also show the approximate numbers of major categories of deaths expected (in thousands) between ages 25 to 69 years (Table 1).

Insert Figure 2 ; Insert Table 1

As of September 2005, 140,000 VA reports have been collected from all SRS units, and about 30,000 records have undergone double physician coding and reconciliation.

Overall we expect several thousand tuberculosis, vascular and cancer deaths among adults. Such numbers are not excessively large, particularly if the age- and sex-specific relevance of several risk factors is to be assessed simultaneously. For example, assuming a power of 90% and two-sided alpha of 0.0001, and assuming a 40% smoking rate among male controls, the study would have sufficient power to detect relative risks among men as low as 1.4 for lung cancer, 1.1 for all cancers or for tuberculosis and 1.1 for cardiovascular disease (28). Thus, the study has robust statistical power to detect small but significant increases in risk for most key variables of interest.

The main planned analyses involve simple tabulations and standard Cox proportional hazards analyses, calculating relative risks that are standardized for age, educational level and selected covariates as relevant. Analyses of adult deaths will focus on deaths at ages 25-69 years as these are much less likely to be misclassified than deaths occurring over age 70.

Sample size estimates for the 2004-2014 sample frame are less certain as the cause-specific child and adult mortality will depend on the rapidity of declines in childhood mortality, and increases in HIV-1 related mortality . However it is reasonable to assume that with expanding sample size (reflecting the growth of populations in the SRS households), some 700,000 deaths will occur over the ten year period.

Reliable measurement of absolute rates and relative risks

The ability to generate absolute rates depends on completeness of enumeration, and being able to calculate underlying demographic denominators, including migration. Evaluations suggest that the SRS has a high ascertainment rate of expected events, although adjustments may be needed for certain age groups. A 1984 study (29) concluded that the system captured 90% of deaths between 1971 and 1980. The SRS, which employs continuous enumeration, is more sensitive at detecting child deaths than are single surveys, and has recorded more child deaths than those estimated by the National Family Health Survey-2 (NHFS-2). Bhat (30) found that adult deaths were underreported by about 13 to 14% (slightly higher in females) and that there is evidence to suggest that the undercount has increased slightly in particular states recently. It is expected that the new SRS sample frame should have corrected the under-reporting of adult deaths noted in the first SRS sample frame. Formal demographic evaluation will be done once results from the new SRS are collected. Because household composition is updated every six months, and in-migration and out migration are traced, it should be possible to calculate absolute mortality rates in the population, including person-years at risk. Deaths among in-migrating groups can be excluded as the death report lists if the person was a usual resident in the SRS unit or not. Some periodic correction for undercount will be needed.

Loss to follow up

The SFMS baseline was confined to usual residents of households only. Those living in these households in 1998 were a sub-selection of the SRS baseline sample frame originally done in 1993. Experience from a prospective study in Chennai (31) suggests that if people are resident

for a few years, they are not likely to move again. Thus, the enrolled group in the current SRS sample frame is less likely to migrate than would be the general population. This is supported by data from our own small pilot examination of 389 randomly chosen SRS records in two northern states, which showed that only 6% (25/389) of the households had moved from 1998 to 2002.

Nonetheless, we do expect loss to follow up from out migration from the SRS unit. The risk ratios would be biased downwards if such loss to follow up is non-differential between exposed and unexposed groups. Similarly, risk ratios would be biased downward if impoverished people (who are most likely exposed) migrate.

Results

Validation studies of mortality outcomes

Verbal autopsy relies on the assumption that most causes of death have distinct symptoms and signs that can be recognized, recalled and reported by household members or associates of the deceased to a trained, (usually) non-medical fieldworker. Further, it is assumed that deaths characterized through VA possess a distinct set of features that can be distinguished from other underlying causes of death (32). Thus, diseases with very distinct symptoms and signs, such as tetanus, that are recognized by the local population may be more suitable for VA than systemic diseases such as malaria, which is associated with many common symptoms and signs. Factors that influence the validity and reliability of VA include the VA instrument (mortality classification, diagnostic procedures), the data collection procedures (recall period, interviewer's characteristics, respondent's characteristics), and the underlying distribution of cause-specific mortality in a given population (33-35). Although there is variation in the sensitivity and specificity for specific conditions, verbal autopsies are now of established value in helping to classify the broad patterns of mortality for childhood deaths in populations that are not covered by adequate medical services. Verbal autopsies have also been used to assess the causes of maternal deaths (36).

Background work for this study included two validation studies of adult deaths in India (37-39). The first study (37,38) developed and tested a VA instrument among 48 000 adult deaths in urban Chennai and 32 000 adult deaths in rural Tamil Nadu, including a 5% random re-sample. VA conducted by trained non-medical fieldworkers resulted in 90% successful reporting on cause of deaths for middle-age adults (25-69 years). The VA instrument reduced the proportion

of adult (age 25 or older) deaths attributed to unspecified or unknown causes from 54% to 23% in urban areas and from 41% to 26 % in rural areas. Verbal autopsy yields fewer unspecified causes (only 10%) than the death certificate (37%), particularly for the deaths that did not occur in hospital, and often yielded somewhat more specific information, e.g., about the approximate site of origin of a cancer, or about evidence of tuberculosis, stroke, myocardial infarction or diabetes (Table 2). The urban Tamil Nadu study also compared VA results to those from a Chennai population-based cancer registry. The VA sensitivity to identify cancer was 95% in the age group 25-69 years and VA identified 288 deaths which were not registered in the Chennai Cancer Registry.

Insert Table 2

The second validation study compared all-cause mortality determined by VA against hospital-based records for 262 adult deaths in northern India (39). Deaths characterized with VA were compared to medically certified cause of death certificates for patients who had died in a hospital. Cause-specific mortality fractions assigned by the verbal autopsy method were statistically similar to the causes arrived at by review of hospital records ($p>0.05$). Specificity was high (>95%) for all broad cause groups except cardiovascular (79%) diseases. Sensitivity was highest for injuries (85%) and it was in the range of 60% to 65% for circulatory diseases, neoplasms, and infectious diseases. Sensitivity was low (20% to 40%) for respiratory, digestive and endocrine diseases. These figures are broadly in agreement with the results from a multicentre validation study of VA for adult deaths conducted in Africa, which found a sensitivity and specificity of 82% and 78%, respectively, for all communicable diseases, and a sensitivity and specificity of 71% and 87%, respectively, for all non-communicable diseases (34).

Preliminary results from two states indicate that the distribution of underlying causes of death based on the random re-interview does not differ substantially from the cause of death derived from the VA reports of the original RGI surveyors (Table 3).

Insert Table 3

Validation studies of exposures measured at baseline

Baseline exposures were captured in a one-time Special Fertility and Mortality Survey (SFMS) conducted within the SRS in February 1998. Baseline exposures captured in the SFMS

included socioeconomic information (education, occupation, household income, household composition), water and sanitation facilities and other living conditions, smoking and its type (bidi, cigarette, hukka or other) and age at onset, alcohol use and frequency per week, past history of various medical conditions, and the type of cooking fuel used (major and second, and use of a separate kitchen). The survey also recorded deaths although not their causes in 1997.

The 2004 baseline survey in the new SRS sample frame (2004-2014) recorded similar exposures and added history of disability from various medical conditions, recent short-term illnesses, tobacco smoking and chewing, alcohol use, vegetarianism, and maternal history including contraceptive use and pregnancies.

The reliability and representativeness of the SFMS baseline survey can be assessed, in part, by comparing the age-specific prevalence of smoking and alcohol consumption among males found in the SFMS and other standard surveys, such as the Indian Census or the National Family Health Survey 2 (NFHS-2; a nationally representative demographic and health survey that interviewed 90,000 women aged 15-49; http://nfhsindia.org/index.html). The supplemental information provides additional validation studies such as birth order of children, and measures of indoor air pollution.

Age-specific prevalence of smoking in adult males: Smoking is currently common only among Indian males. The SFMS and NFHS-2 report very similar trends in age-specific prevalence of current male smokers. The steepest increase in prevalence of smoking is between 20-30 years of age (Figure 3). Similarly, the prevalence of smoking among males across 25 states shows a strong correlation between the two studies (Pearson correlation coefficient or R^2 of 0.92; data not shown). We plan to re-measure the extent of smoking by each of several thousand living adults so as to correct for regression-dilution bias (13). Additionally, to verify and validate smoking status, we will use a simple hand held carbon monoxide breathalyzer on about 1000 smokers and non-smokers in select states. These simple breathalyzers appear to be effective at detecting smoking status (40).

Insert Figure 3

Age-specific alcohol consumption in adult males: As with smoking, a similar pattern in age specific alcohol consumption is captured by the SFMS and NFHS-2 (Figure 4). The

discrepancy in absolute prevalence for each age group may be due to the difference in the sex of the respondents, where females are either over reporting male alcohol consumption in the household in the NFHS-2, or males are under-reporting their own consumption in the SFMS. It should also be noted that, unlike the NFHS-2, usually the male head of the household is the respondent of the SFMS.

Insert Figure 4

Retrospective "proportional mortality" and "household case-control" methods

The determinants of death can be identified by comparing risk factors between the dead and living. Such "household case control" studies use the dead as cases and their surviving spouses or close relatives as controls. A retrospective study in Chennai of 43,000 male deaths and 35,000 living controls (41) using these methods has documented that throughout middle age, the death rates from medical causes of smokers were double those of non-smokers (standardized risk ratio at ages 25-69 of 2.1 with 95% confidence interval 2.0-2.2, smoking-attributable fraction 31%). A large part of this excess risk was from tuberculosis and vascular deaths. If these hazards are similar across India, then about half of all tuberculosis deaths in India could be accounted for by smoking. We will apply similar retrospective methods for smoking, tobacco chewing and alcohol as these exposures are gathered for dead adults and from a living household respondent.

Simply asking about the dead person's risk factors could be useful. A recent retrospective study of 1 million deaths in China, compared the proportions of smokers and non-smokers who have died of tobacco-attributable diseases versus non-tobacco related diseases (chiefly injuries), to calculate the excess in smokers (42). Our preliminary pilot studies among childhood deaths in Northern India found that 61% of children (16/26) who died of vaccine-preventable diseases were not immunized in comparison to 40% of control children (18/45) who died from injuries. The crude odds ratio of 2.4 suggests that half of the vaccine-preventable child deaths need not have occurred (data not shown).

Using these two retrospective methods, we are conducting validation studies of selected risk factors such as childhood immunization, childhood malnutrition, alcohol, male time away from home (as a proxy of HIV-1 related sexual risk taking), and other variables.

Discussion

Key design challenges

Several challenges stand out in the design and implementation of this massive study. First is its sheer size. A total of 14 million people is a large sample frame, but commensurate with the needs of monitoring health status among the one billion people in India. The study builds upon the routine infrastructure that has been in place for over 30 years within the SRS. It works with the leading demographic research organization in India (for example, the decennial census involves the RGI hiring two million enumerators to survey about 150 million households within 25 days). Second is the need for simplification of methods to make such scale of study practicable. The study is not the only epidemiological study needed in India, but its unique focus and scale raise important challenges in the development of new methods. The Million Death study uses a "large simple" design, which places specific demands on ensuring that new procedures (such as the new VA instrument and physical and biological measurements) are rigorously tested, piloted, and simplified. The third challenge is sustainable funding. The current SRS sample frame is being followed on a total incremental budget of $2 million US (or less than 33 cents per person). This does not include the core funding of the RGI surveyors and related infrastructure. With inclusion of those, the overall study can still be done in a highly cost-effective manner at well under $1 per person. Collection, processing and long-term storage of biological samples is expected to cost $2-5 per person for dried blood spots, and $10-20 per person for a 10 ml tube of blood. These costs compare extremely favorably with those existing bio-banks in the UK and elsewhere (43;44). Most of the infrastructure costs will be sustained by the Government of India. However, we believe that redirecting some of the considerable (and often inefficient) spending on monitoring individual disease projects is required to enhance the SRS.

Key design issues for blood-based genetic epidemiology

Discovery of "new" risk factors should benefit from the recent and extraordinarily rapid progress by many different research groups and biotechnology companies in developing low-cost, miniaturized methods for the simultaneous assay in small volumes of blood (or, perhaps, dried blood spots) of vast numbers of nucleic acid fragments, host genetic factors, proteins, small molecules and pathogens (45). This rapid biotechnological progress has been backed by increasingly sophisticated computer software to help interpret the mass of numerical information

that can be generated from each person's blood, yielding within the next few years many, as yet unforeseen, qualitatively different analytic capacities. Appropriately large-scale epidemiological studies that acquire blood (or other) samples from individuals and systematically link them to relevant measures of disability and future mortality are required to make such technological progress relevant to human populations. It is particularly important for such studies to address the shortcomings of the existing biological sample collections that are now being undertaken (43;44) including the particular circumstances of India and the specific infectious diseases common to developing countries.

Key issues which arise in moving to blood-based epidemiology include the choice of blood sample, long-term storage and retrieval, and statistical design issues,

Choice of blood sample

Biological specimen collection procedures will build on experience in China, India, and elsewhere. We will undertake pilot studies to test the feasibility and acceptability of methods, and to ensure that biological samples collected can serve as a long-term source of DNA for genotyping, and material for biochemical, hematological, proteonomic and other assays (46). Major options include a 10 ml non-fasting blood sample collected into one 10 ml EDTA vacutainer, which has been shown previously by the University of Oxford laboratory to allow a wide range of assays (47). This system is used by the ongoing Chinese Kadoorie Study of 500,000 adults (48) and by the UK Biobank project of the same size (43). Alternatively, dried blood samples on filter paper have the advantages of easy storage and transport, as well as being less intrusive (i.e., by finger-prick rather than by venepuncture). Dried blood spots have been recommended by the WHO for use in field HIV-1 investigations, and have been used in various studies within India (49;50).

We currently plan pilot studies of dried blood spots or tubes of blood among 4,000 adults in eight SRS units in 4-5 states plus a special survey of 5000 adults in one state.. These pilots will focus on standardizing methods to obtain anthropometric, behavioral and physiologic measurements and to evaluate the alternative methods for collection of biological specimens for their utility, cost and practicability (tube of blood, dried blood spot, urine). The pilots will focus on feasibility of field methods, simplification of approaches, and quality control. The results of the pilot studies, which are expected by June 2006, will help inform the design of the larger survey. The special survey will help define practicable questions on HIV-1 risk behaviors.

Long-term storage, retrieval and analyses

We are undertaking systematic reviews of the literature and reviewing available assays in India to examine which current bioanalytic methods can be used to test current hypotheses (either for correlates of infection or chronic disease). Using the above pilot samples, we will develop testing and analytic methods for biological samples that would permit high-throughput, low-cost and high quality assays to be run, and to permit the long term reliable storage and retrieval of such samples. Dried blood spots may well be stable in a basic 4° C refrigerator, with minimal storage requirements (46). With a population of one billion, India certainly requires at least 1-2 major bio-repositories (including splitting samples, which safeguards against loss at one facility). We are developing plans for long-term bio-repositories with the ability to store samples for decades. Much will depend on the choice of the final assay.

Nested case-control methods and genetic association studies

There is a long list of biological factors in blood that might be correlates of disease-specific mortality. To be efficient, we will use a "nested" case-control approach in our prospective study. Biological samples are taken from all adults in a baseline survey, and stored long-term; then, when sufficient numbers of cases with the disease of interest have died (based on the six month follow up of the causes of death), aliquots from those cases are retrieved from storage, plus aliquots from a few matched controls per case; and, finally, the factors of interest are assayed in these cases and these controls. This design is similar to that in the Chinese Kadoorie study and the UK BioBank project (43). As noted above, the number of deaths is likely to be substantial for most of the common diseases of interest to make this an efficient strategy.

Under some circumstances, studies that used population-based controls to study gene-disease associations have been biased due to underlying variation in gene frequencies between populations ("population stratification"; (51)). Genetic epidemiology can provide reliable population-based estimates of disease-allele frequency, penetrance and attributable risk, particularly if designs that account for this bias are employed, and accordingly this has led to greater emphasis recently on family-based association designs. However, there is also evidence that well-designed population-based studies are sometimes superior (52). The SRS offers a unique opportunity to explore statistical design issues, as both general population and family based sampling is possible. Family-based studies allow for population-specific estimates of clustering of disease and correlation (heritability). Further, families with a large number of

siblings could be over sampled to increase power for genetic linkage studies. The use of 'genomic controls' where anonymous markers are genotyped to test for population stratification can be employed (53).

Limitations

The study has several limitations. First, despite a very large sample size, the statistical power remains limited, especially in a prospective design for less common causes such as those leading to maternal deaths. However, for most of the major public health conditions of importance, sufficient events should occur to generate plausible absolute rates and relative risks for risk factors. Second, careful attention in design means that most identified biases should be minimized, and should ensure high internal consistency of field work and coding. However, periodic revalidation of mortality outcomes against external standards will be needed. Similarly, continuous improvement of exposure measurements, partially through careful pilots, will be needed. Third, VA yields broad classification of the underlying causes in about 90% of deaths before age 70. In old age, however, the proportion of classifiable deaths is substantially lower. For some specific conditions, such as childhood pneumonia, the cause-specific fractions may be difficult to estimate below certain levels (54).

Significance to global health

This study will reliably document not only the underlying cause of child and adult deaths, but key risk factors (behavioral, physical, environmental and eventually, genetic). It offers a globally-replicable model for reliably estimating cause-specific mortality using VA, and strengthens India's flagship mortality monitoring system. Despite the misclassification that is still expected, the new cause of death data will be a quantum better than that available previously in India. It builds India's capacity for research and for public health action. It provides a large, representative, low-cost and long-term system to reliably track the health status of one billion people for the next decade or longer.

Contributors
P Jha initiated the study and all authors on the writing committee contributed to the implementation of the validation studies as well as the design and writing of this study protocol.

RGI-CGHR Prospective Study Collaborators:

RGI-CGHR National Centre, Office of the Registrar General, RK Puram, New Delhi, India: DK Sikri*, RC Sethi* +, N Dhingra * +, DK Dey, M Jain, S Jain, K Lal, L Sushant

Indian Academic Partners:
Clinical Epidemiology Resource and Training Centre, Trivandarum: KB Leena, KT Shenoy*
Department of Community Medicine, Gujarat Medical College, Ahmedabad: DV Bala, P Seth, KN Trivedi*
Department of Community Medicine, Kolkatta Medical College, Kolkatta: SK Roy*
Department of Community Medicine, Osmania Medical College, Hyderabad: P Bhatia*
Department of Community Medicine, Regional Institute of Medical Sciences, Imphal: L Usharani*
Department of Community Medicine, SMS Medical College, Jaipur: AK Bharadwaj*
Epidemiological Research Centre, Chennai: V Gajalakshmi* +
Gandhi Medical College, Bhopal: R Dikshit*, S Sorangi
Healis-Seskarhia Institute of Public Health, Navi Mumbai: PC Gupta* +, MS Pednekar, S Sreevidya
Institute of Population Health and Clinical Research, St. John's Medical College, Bangalore: A Kurpad, P Mony* +, M Vaz
King George Medical College, Lucknow: S Awasthi*
North Eastern Indira Gandhi Institute of Regional Medical Sciences, Shillong, Meghalaya: FU Ahmed*
Regional Medical Research Center, ICMR Institute, Bhubaneshwar: AS Karketta*, K Dar
School of Preventive Oncology, Patna: DN Sinha*
School of Public Health, Post Graduate Institute of Medical Education and Research, Chandigarh: N Kaur, R Kumar* +, JS Thakur

Other Partners:
Clinical Trial and Epidemiological Studies Unit, University of Oxford, Oxford, England: Z. Chen, R. Collins, Sir R Peto * +
Hospital for Sick Children, University of Toronto, Toronto, Canada: A Patterson, S Schrier
Indian Council of Medical Research, New Delhi, India: NK Ganguly*
Mt. Sinai Hospital, University of Toronto, Toronto, Canada: J McLaughlin
McLaughlin Centre for Molecular Medicine, University of Toronto, Toronto, Canada: K Kain, R Kaul
Royal Netherlands Tuberculosis Program, The Hague, Netherlands: N Naglekerke
World Health Organization, Geneva, Switzerland: T Boerma*, T Evans*, K Shibuyi
World Health Organization, South East Asia Regional Office, New Delhi, India: N Singh, T Sein

Global Coordinating Centre, Centre for Global Health Research, St. Michael's Hospital, University of Toronto, Canada: B Jacob, P Jha (Principal Investigator)*+, R Kadmood, C Major, J Moore+, P Parra, S Sgaier, H Shadmand, BL Shi, D Thiruchelvam, P Vasa+, F Zhang

** Member Advisory Committee*
+ Writing committee for this report

Role of the funding source

The sponsors of the study had no role in the study design, data collection, data analysis, data interpretation or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility in the decision to submit for publication.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgements

The largest proportion of the study costs are met by the Government of India as part of the routine costs of running the SRS. External funding from the study comes from the National Institute of Health Tobacco Research Grant (R01 TW05991-01; Aron Primack), the Canadian Immunization Initiative of the International Developmental and Research Centre (Grant no 102172; Sharmila Mhatre), Canadian Institute of Health Research (Establishment Grant No IEG-53506; Mark Bisby), and unrestricted grants from the McLaughlin Centre for Molecular Medicine, University of Toronto (Duncan Stewart), St. Michael's Hospital (Arthur Slutsky) and University of Toronto (Harvey Skinner).

Prabhat Jha is supported by a Canada Research Chair of the Government of Canada.

We thank Dr. Paul Doherty for editorial assistance.

Reference List

(1) World Health Organization (2002) Reducing Risks: Promoting Healthy Life: World Health Report. Geneva, Switzerland: World Health Organization.

(2) Editorial. (2005) Stumbling around in the dark. Lancet 365:1983.

(3) Horton R. (2005) The Ellison Institute: Monitoring health, challenging WHO. Lancet 366:179-181.

(4) Stansfield S (2005) Structuring information and incentives to improve health. Bulletin of WHO 83: 562-563.

(5) Murray CJ, Lopez AD, Wibulpolprasert S (2004) Monitoring global health: time for new solutions. BMJ 329:1096-1100.

(6) Mitra B. (1999) India's mortality measurement systems. In: Centers for Disease Control and Prevention. Counting the dead in India in the 21st century. S. Asma, P. Jha, PC Gupta eds. Proceedings of the International Workshop on Certification on Causes of Death, Mumbai. US Centers for Disease Control.

(7) Jha P (2001) Reliable Mortality Data: A Powerful Tool for Public Health. National Medical Journal of India 14:129-131.

(8) Jha P, Slutsky AS, Brown D, Nagelkerke N, Brunham BG, Bergeron MG et al. Global IDEA Scientific Advisory Committee (2004) Health and economic benefits of an accelerated program of research to combat global infectious diseases. CMAJ 171:1203-1208.

(9) Doll R, Peto R, Boreham J, Sutherland I (2004) Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ 328: 1519.

(10) Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, et al. (1981) Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N Eng J Med 305:1425-143l.

(11) Jha P (2002) Avoidable mortality in India: past progress and future prospects. Natl Med J India 15(Suppl 1):32-36.

(12) Lewington S, Clarke R, Qizilbash N, Peto R, Collins R (2002) Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet 360: 1903-1913.

(13) Clarke R, Shippley M, Lewington S, Youngman L, Collins R, Marmot M, et al. (1999) Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. Am J Epidemiol 150:341-353.

(14) Yusuf S, Hawken S, Ounpuu S, et al and INTERHEART Study Investigators (2004) Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 364: 937-952.

(15) Danesh J, Collins R, Peto R (2000) Lipoprotein (a) and coronary heart disease: meta-analyses of prospective studies. Circulation 102:1082-1085.

(16) Knoblauch H, Bauerfeind A, Toliat M, Becker C, Luganskaja C et al. (2004) Haplotypes and SNPs in 13 lipid-relevant genes explain most of the genetic variance in high-density lipoprotein and low-density lipoprotein cholesterol. Hum Mol Genet 13: 993-1004.

(17) Gonzalez E, Bamshad M, Sato N, Mummidi S, Dhanda R, et al. (1999) Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. Proc Natl Acad Sci USA 96:1204-1209.

(18) Shanmugalakshmi S, Pitchappan R (2002) Genetic basis of tuberculosis susceptibility in India. Indian J Pediatr 69(Suppl 1): S25-S28.

(19) MacDonald K, Fowke K, Kimani J, Dunand VA, Nagelkerke NJ, et al. (2000) Influence of HLA supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection. J Infect Dis 181:1581-1519.

(20) Kaul R, Kimani J, Nagelkerke NJ, Fonck K, Ngugi EN, et al. (2004) Monthly antibiotic chemoprophylaxis and incidence of sexually transmitted infections and HIV-1 infection in Kenyan sex workers: a randomized controlled trial. JAMA 291: 2555-2562.

(21) Nagelkerke NJ, de Vlas SJ, MacDonald KS, Rieder HL (2004) Tuberculosis and sexually transmitted infections. Emerg Infect Dis 10: 2055-2056.

(22) WHO. (2003) Manual of International Classification of Diseases, Injuries and Causes of Death (Tenth revision). Geneva, Switzerland: World Health Organization.

(23) Banthia J, Dyson T (1999) Smallpox in 19th Century India. Population and Development Review 24: 649-680.

(24) Caselli G (1991) Health Transition and Cause-specific Mortality. In The Decline of Mortality in Europe. Schofield R, Reher D, Bideau A, Eds. Oxford Clarendon Press.

(25) Gupta PC, Sankaranarayanan R, Ferlay J (1994) Cancer deaths in India: is the model-based approach valid? Bull World Health Organ. 72: 943-944.

(26) Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC (1999) Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA 282:677-686.

(27) Registrar General (2001) Compendium of India's Fertility and Mortality Indicators 1971-1999.

(28) Schlesselman S (1982) Case-control Studies: Design, Conduct, Analysis. New York: Oxford U. Press.

(29) Preston S, Bhat P (1984) New evidence on fertility and mortality trends in India. Population and Development Review 10: 481-503.

(30) Bhat PN (2003) Completeness of India's sample registration system: an assessment using the general growth balance method. Popul Stud (Camb) 56: 119-134.

(31) Gajalakshmi V. (2005) Personal Communication

(32) Anker M, Black R, Coldham C, Kalter H, Quigley M, et al. (1999) A standard verbal autopsy for investigating causes of death in infants and children. WHO; Report No.: WHO/CDS/CRS/ISR/99.4.

(33) Quigley MA, Chandramohan D, Rodrigues LC (1999) Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. Int J Epidemiol 28:1081-1087.

(34) Chandramohan D, Maude GH, Rodrigues LC, Hayes RJ (1998) Verbal autopsies for adult deaths: their development and validation in a multicentre study. Trop Med Int Health 3: 436-446.

(35) Kalter HD, Gray RH, Black RE, Gultiano SA 1991 Validation of the diagnosis of childhood morbidity using maternal health interviews. Int J Epidemiol 20: 193-198.

(36) Kumar R, Sharma AK, Barik S, Kumar V (1989) Maternal mortality inquiry in a rural community of north India. Int J Gynaecol Obstet;29: 313-319.

(37) Gajalakshmi V, Richard P, Santhanakrishnan K, Sivagurunathan B (2002) Verbal autopsy of 48,000 adult deaths attributed to medical causes in Chennai (formerly Madras), India. BMC Public Health 2:7.

(38) Gajalakshmi V, Peto R (2004) Verbal autopsy of 80,000 adult deaths in Tamil Nadu, South India. BMC Public Health 4: 47.

(39) Kumar R, Thakur J, Rao M, Singh M, Bhatia P (2005) Validity of Verbal Autopsy in Determining Causes of Adult Deaths. Indian Journal of Public Health. In press.

(40) Cunnington AJ, Hormbrey P (2002) Breath analysis to detect recent exposure to carbon monoxide. Postgrad Med J 78: 233-237.

(41) Gajalakshmi V, Peto R, Kanaka TS, Jha P (2003) Smoking and mortality from tuberculosis and other diseases in India: retrospective study of 43000 adult male deaths and 35000 controls. Lancet 362: 507-515.

(42) Liu BQ, Peto R, Chen ZM, Boreham J, Wu YP, Li JY, et al. (1998) Emerging tobacco hazards in China: 1. Retrospective proportional mortality study of one million deaths. BMJ 317:1411-1422.

(43) Biobank Project, 2004. http://www.biobank.ac.uk. Accessed May 8, 2005.

(44) Kaiser J (2002) Biobanks. Population databases boom, from Iceland to the U.S. Science; 298: 1158-1161.

(45) Varmus H (2003) Genomic empowerment: the importance of public databases. Nat Genet 35(Suppl 1): 3.

(46) Steinberg K, Beck J, Nickerson D, Garcia-Closas M, Gallagher M, Caggana M, et al. (2002) DNA banking for epidemiologic studies: a review of current practices. Epidemiology 13: 246-254.

(47) Clark S, Youngman LD, Palmer A, Parish S, Peto R, Collins R (2003) Stability of plasma analytes after delayed separation of whole blood: implications for epidemiological studies. Int J Epidemiol 32:125-130.

(48) Chen Z (2005). Personal Communication.

(49) Solomon SS, Solomon S, Rodriguez II, McGarvey ST, Ganesh AK, Thyagarajan SP, et al. (2002) Dried blood spots (DBS): a valuable tool for HIV surveillance in developing/tropical countries. Int J STD AIDS 13: 25-28.

(50) Ramakrishnan L, Reddy KS, Jailkhani BL (2001) Measurement of cholesterol and triglycerides in dried serum and the effect of storage. Clin Chem 47:1113-1115.

(51) Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489-493.

(52) Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598-604.

(53) Devlin B, Roeder K (2000) Genomic control for association studies. Biometrics 55: 997-1004.

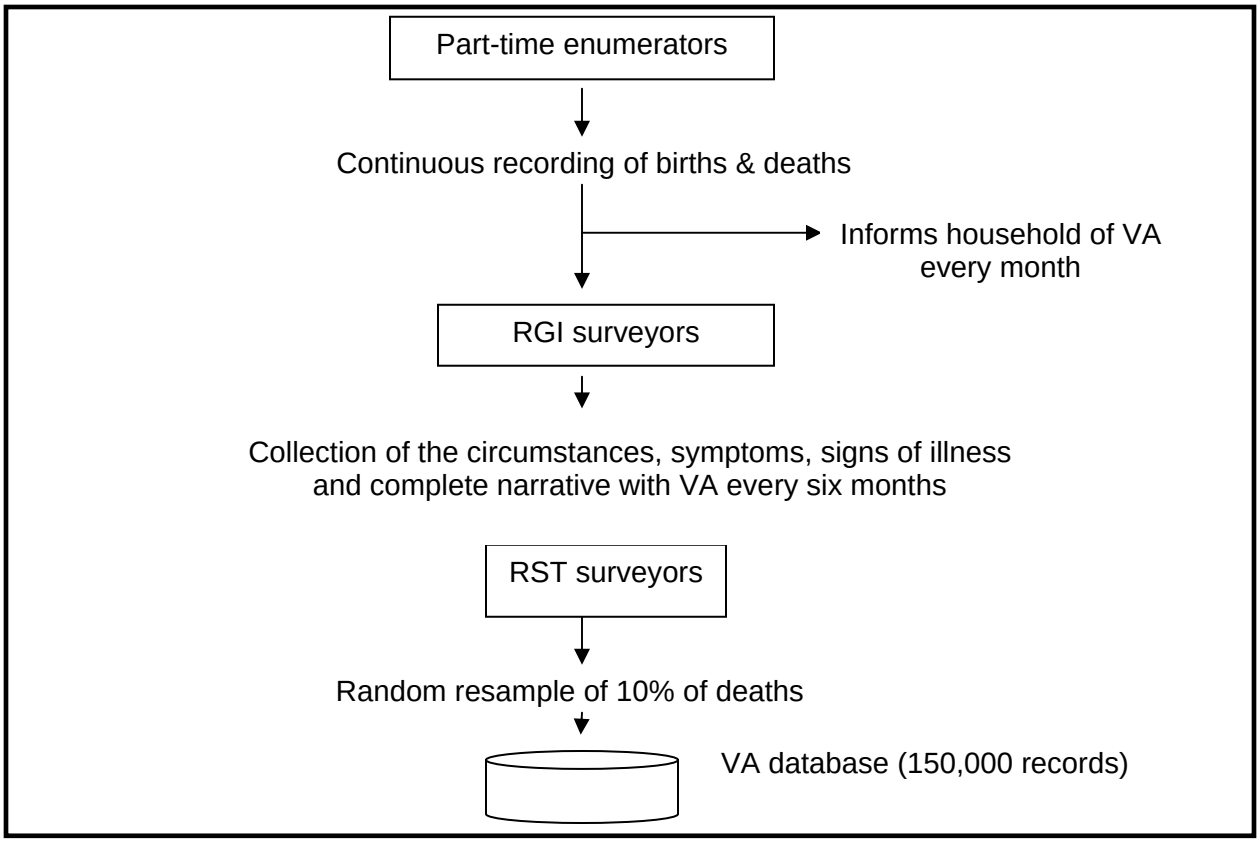
(54) Williams BG, Gouws E, Boschi-Pinto C, Bryce J, Dye C (2002) Estimates of world-wide distribution of child deaths from acute respiratory infections. Lancet Infect Dis. 2: 25-32.

Box 1. SRS verbal autopsy methods overview

Design of verbal autopsy questionnaire	*Combined open/closed format. Structured questions accompanied by an open-ended narrative. Symptom list to assist attribution of deaths.*
Questionnaire layout	*One-page, double-sided, scannable forms. Four age-specific forms (neonatal, child, adult and maternal). Forms available in either English or Hindi.*
Interviewers	*Non-medical RGI surveyors (mostly male) with knowledge of local language(s) and trained in VA instrument.*
Interview technique	*One-on-one interviews during home visits. Duration of 30-45 mins.*
Respondents	*Family members or other informants (usually neighbors or close associates of the deceased).*
Recall period	*Usually <six months, but useful up to three years*
Data quality	*Random resample of 10% of all deaths to ensure completeness of fieldwork.*
Derivation of diagnosis	*Central medical review of cause by two independent physicians using modified VA reports (Physician Reports) and an internet-based web application. Adjudication of disagreements by an expert physician.*
Mortality classification	*The International Classification of Diseases, 10th Version (ICD-10)*
Sample size	*1 million deaths (about 0.3 M from 1998-2003, 0.7 M from 2004-2014)*

Figure 1. SRS verbal autopsy activities

A. Field data collection among 1.3 million households



B. Cause of death assignment

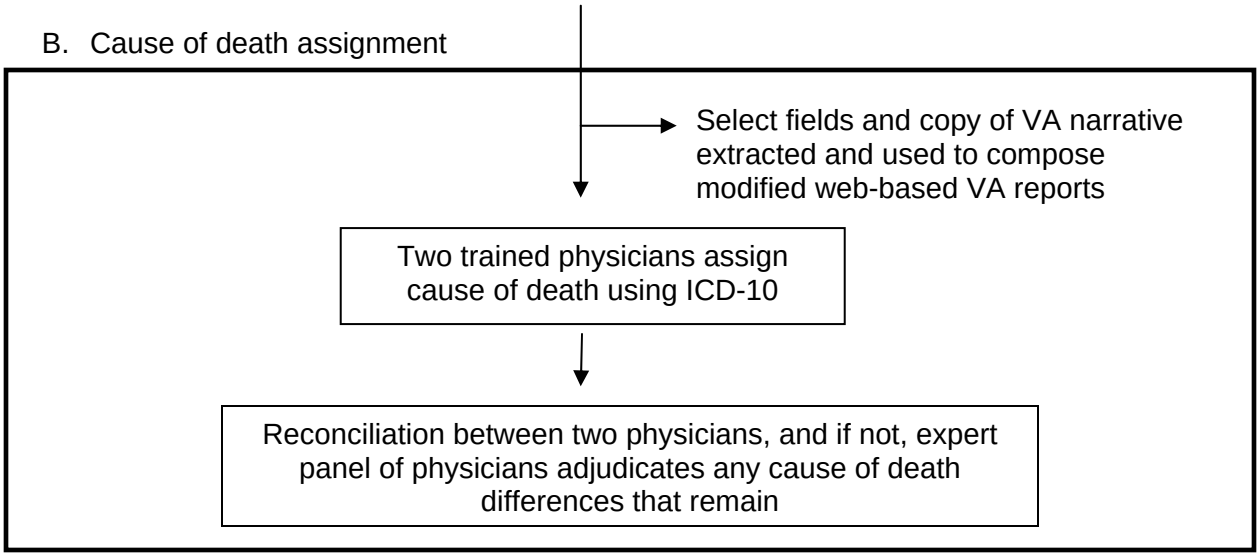


Figure 2. Expected number of deaths by age group using verbal autopsy ('000s), 2001 to 2003

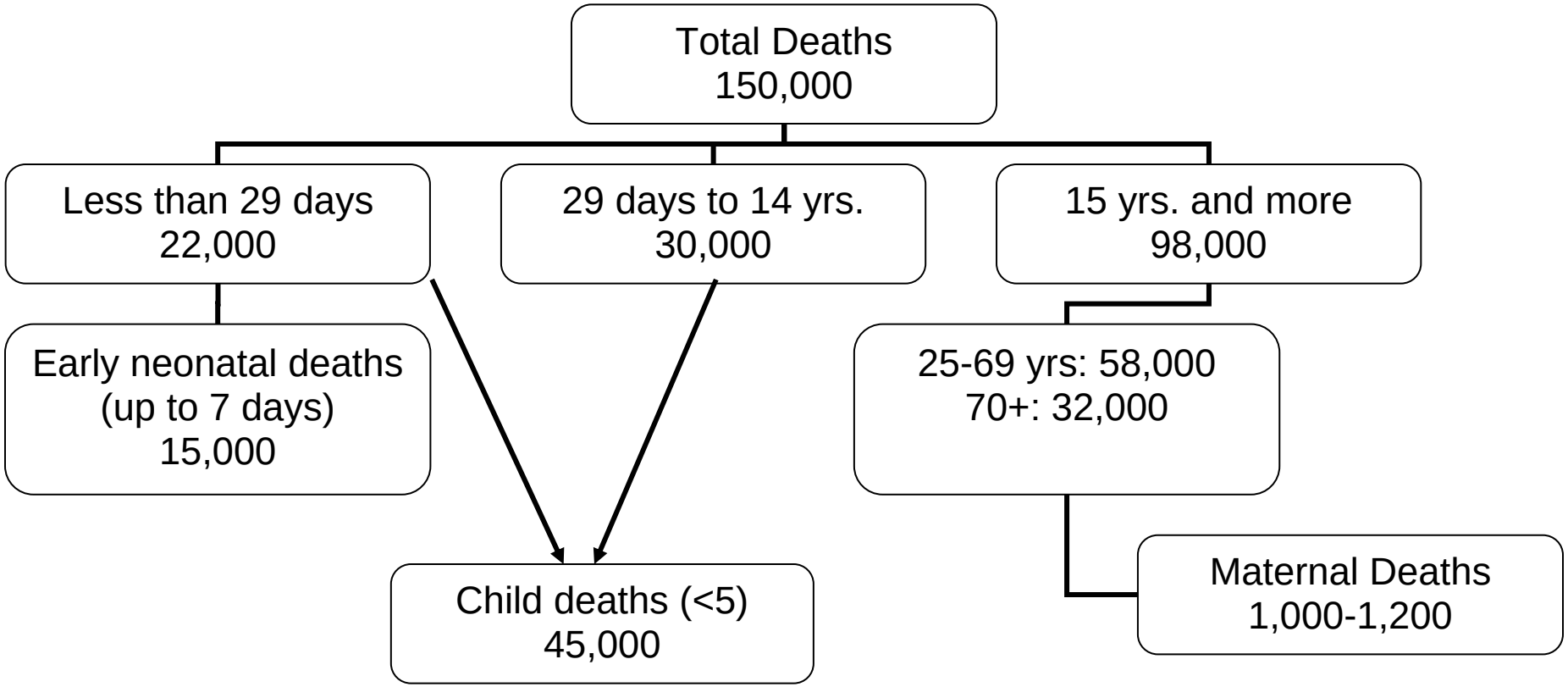


Figure 3. Prevalence of current male smokers, Special Fertility and Mortality Survey (SFMS) 1998 vs. National Family Health Survey (NFHS-2) 1998-1999

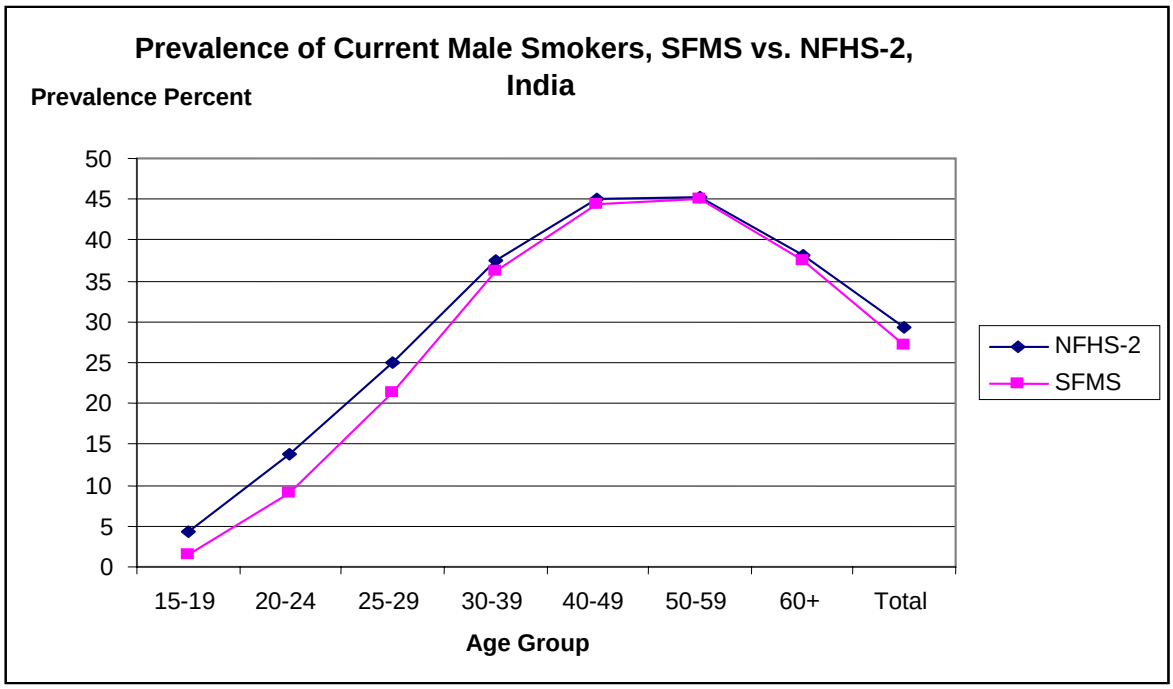


Figure 4: Prevalence of current male alcohol drinkers, Special Fertility and Mortality Survey (SFMS) 1998 vs. National Family Health Survey (NFHS-2) 1998-1999

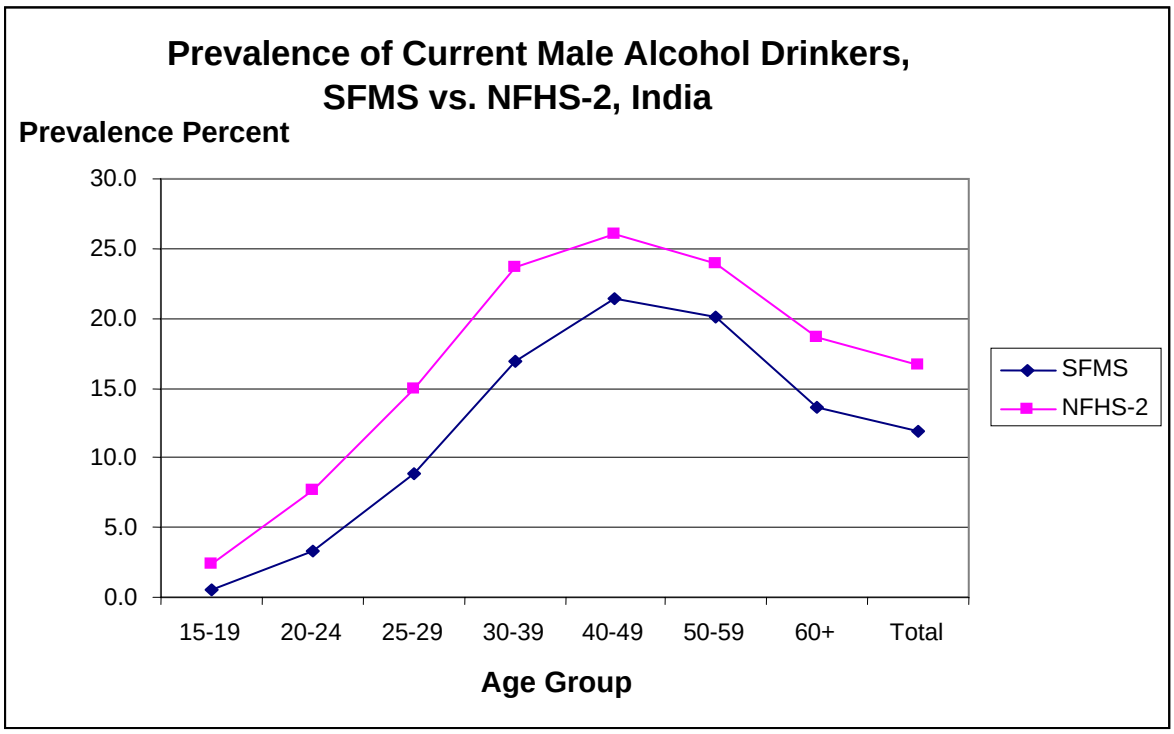


Table 1. Expected number of deaths between ages 25 and 69 by cause ('000s), 2001 to 2003

Causes	Male	Female
All infections and maternal diseases	9.2	5.9
Tuberculosis	*3.4*	*1.5*
HIV/AIDS	*1.9*	*0.4*
Lower respiratory infections	*1.3*	*1.0*
Maternal conditions	*N/A*	*1.1*
Other infections	*2.6*	*1.9*
All non-communicable diseases	21.2	14.6
Lung cancer	*0.8*	*0.2*
Upper aerodigestive cancer (mouth, oropharynx, oesophagus)	*0.8*	*0.3*
Other cancers	*1.8*	*2.3*
Cardiovascular diseases	*11.4*	*7.6*
Chronic obstructive pulmonary disease	*2.4*	*1.5*
Other non-communicable	*4.0*	*2.7*
Injuries	4.5	2.5
Unintentional	*3.2*	*1.9*
Intentional	*1.3*	*0.6*
Total (000's)	**34.9**	**23.0**

Note: The expected numbers of deaths for these causes were calculated by using the 2001 estimates of mortality from the global burden of disease for India (1; and www.fic.nih.gov/dcpp/gbd.html) and multiplying the proportion of specific causes by the total age-specific number of deaths expected in the SRS study during the follow-up from 2001 to 2003.

Table 2: Cause of death based on vital statistics and verbal autopsy of 48,000 adult (> 25 years of age) deaths in Chennai, India: 1995-97; (Reference 33-34)

Cause of Death	**Cause of death in vital statistics division**		**Cause of death based on verbal autopsy**	
	Male (%)	**Female (%)**	**Male (%)**	**Female (%)**
Vascular disease	8,319 (30)	5,168 (25)	11,056 (41)	7,435 (37)
Tuberculosis (TB)	1,399 (5)	372 (2)	2,231 (8)	575 (3)
Other respiratory diseases	1,088 (4)	596 (3)	1,597 (6)	855 (4)
Neoplasm	1,163 (4)	1,002 (5)	2,344 (9)	1,999 (10)
Infection (excluding respiratory and TB)	584 (2)	303 (2)	1,034 (4)	618 (3)
Unspecified medical	12,291 (44)	11,511 (56)	4,367 (16)	5,889 (29)
Other specified medical	1,899 (7)	1,045 (5)	4,414 (16)	2,804 (14)
Cause not known	983 (4)	634 (3)	Nil	Nil
Total deaths - medical	27,726	20,631	27,043	20,175
External causes	Excluded		683	456
Total deaths (medical + external)	27,726	20,631	27,726	20,631

Table 3. Cause of death results for RGI surveyors and resample teams in Tamil Nadu and Maharashtra and Goa using verbal autopsy (deaths > 28 days) in 2003

Causes	Tamil Nadu				Maharashtra/Goa			
	RGI surveyors		Re-sample team		RGI surveyors		Re-sample team	
	No.	%	No.	%	No.	%	No.	%
Vascular	300	25.8	46	28.2	412	23.5	80	25.0
Tuberculosis (respiratory)	50	4.3	1	0.6	85	4.8	8	2.5
Other respiratory diseases	105	9.0	10	6.1	154	8.8	22	6.9
Cancer	100	8.6	18	11.0	113	6.4	15	4.7
Infection	127	10.9	24	14.7	139	7.9	44	13.8
Diabetes	42	3.6	3	1.8	56	3.2	6	1.9
Peptic ulcer	21	1.8	0	0	7	0.4	2	0.6
Undefined/inadequate information	164	14.1	23	14.1	152	8.7	45	14.1
Other specified	95	8.2	12	7.4	480	27.4	74	23.1
External causes	161	13.8	26	16	155	8.8	24	7.5
Total (n)	**1,165**	**-**	**163**	**-**	**1,753**		**320**	

Supplemental Text: Additional details of field collection methods, study organization and validation experience to date

I. Field Methods

Data collection with the VA instrument

VA forms were designed after seeking expert opinion from the World Health Organization (WHO; (1)), and after review of the literature on validation studies (1-4). The methodology of writing the VA report is based on the large-scale studies done in Tamil Nadu (4;5). The first VA forms were extensively piloted in five states in April 2002, and their results were reviewed by an expert panel of WHO scientists in June 2002. The early review found that even basic training (one day) decreased the proportion of deaths for which there was only a one-word narrative from 50% to 5%. With more substantial training, the RGI surveyors were able to raise the proportion of deaths for which a cause is available from about 50% to over 85% of deaths before age 70.

Based on these pilots and review, we decided to adopt and develop an open/closed format (forms are found at www.cghr.org/project.htm). We conducted a small pilot of 262 adult deaths and found that if the open-ended narrative alone is used, trained physician coders were able to assign an underlying cause of death for 56% of the records (6). If the questions alone are provided, then trained physician coders were able to assign an underlying cause of death for only 49% of the VA records. When the narrative and questions were provided together, the proportion of classifiable VA records rose to 65%. Our results suggest that high quality narratives are the single most important factor in increasing the proportion of classifiable causes of death.

In the closed section, “filter” questions are used to elicit the presence or absence of specific signs and symptoms. If the filter question is positive, then subsequent questions are provided to determine the severity, duration, or other characteristics of these symptoms. For adult deaths, there is a list of signs and symptoms used by the RGI surveyor to obtain more detailed information about the cause of death. The symptom list is used as a filter to define additional probing questions that should be asked if the respondent mentions a particular symptom during the verbatim account of the death. RGI surveyors go through each symptom to ensure that key filter symptoms have not been missed. The written narrative details the following information: associated signs

and symptoms in chronological order; duration; sudden or gradual onset of illness; type of treatment if any treatment received; details on hospitalization prior to death; name and location of hospital; duration of hospitalization; history of similar episodes and treatment given; and abstract information related to the illness prior to death from available investigation reports, death certificate, or discharge summaries. Each interview lasts about 30-45 minutes.

Among the first few thousand deaths, the proportion of VA records coded with high certainty of diagnosis does not vary according to relationship of respondent to the head of the household (family relative, other relatives or neighbor; data not shown). Whether or not the respondent lived with the deceased during the illness that lead to death is the more important determinant of obtaining classifiable causes.

All forms have a common format that includes a socioeconomic and demographic profile of the respondent and the deceased, details of the illness and a narrative section. Forms are available in English or Hindi, with the narrative written in the local language. Our current VA instrument is in its seventh generation and is provided as a single-page, double-sided layout in easy-to-carry booklets with simple instructions for their use. Each type of form (neonatal, child, adult, and maternal) is color-coded and bound according to the type of form for ease of identification in the field.

Random re-sample of RGI surveys

To ensure high quality fieldwork, a specialist re-sample team directly reporting to the study investigators re-interviews 10% (randomly chosen) of households. These are also submitted for central medical review and revalidated. During the early phase of the study, about 10% of each RGI surveyor's visits will be re-sampled, so as to provide early training input and correction of methods. During the later phases, the percentage will be reduced to about 5%. On average, each state will have about two re-sample team members, each covering about one unit per RGI surveyor (i.e., about 1 out of the 10-12 units sampled by a RGI surveyor), or about 15 units per re-sample team. RGI surveyors will receive feedback on the completeness of their work from the work of the re-sample team. As noted in the main text (table 2), the correlation between the random audit team and the RGI supervisors on overall distribution of causes of death was high.

Double coding by trained physicians

Previous validation results of VA for adult deaths suggested that central diagnosis by a trained panel of physician coders yielded consistently higher sensitivity for the cause of most specific mortality outcomes than opinion-based algorithms (7). For child deaths, physician coding is comparable or better than algorithms (8). In this study, in order to reduce inter-observer variation, two trained physician coders independently examine each VA report and determine a probable underlying cause of death coded in the International Classification of Diseases-10th revision (ICD-10; (9)). Before assigning cause of death, the physician coders are trained to carefully screen all relevant information provided, noting all of the positive evidence, and use clinical judgment in assigning the underlying cause of death. For each VA record, physician coders will provide the following information: an underlying cause of death in words (e.g., "tuberculosis"); corresponding ICD-10 code (e.g, A15); and the key words used to guide and support their decision. If two physicians do not agree on an underlying cause of death, a web-based system assigns to each physician the original report and the ICD-10 code of the other physician (without revealing the identity of the other physician). The physician coders are then required to use the additional information provided by the third physician (ICD-10 and key words) to reach an agreement on underlying cause of death. An expert panel of senior physician coders will review VA records where two physicians cannot agree on a cause of death after one reconciliation attempt. Physicians are drawn from across India, so as to ensure that cross-state comparisons are valid.

A pilot of double coding of 1,198 VA records was conducted in the Karnataka state. In 84% of the cases, two independent physicians were able to code to a common cause after only one round of VA training. We expect approximately half of the outstanding differences to be "minor" differences that should be easily resolved after a reconciliation attempt, thus yielding about 85 to 90% first-round agreements between two independent physician coders.

Pilot studies of physical, behavioral and biological measurements

Physical measurements for adults will involve blood pressure, height, weight, waist/hip circumference, and lung function. From children, simple height and weight data will be collected. We place special emphasis on obtaining measures of adult obesity, especially as the patterns differ greatly from developed countries. Several of our collaborators have

implemented simple physical measurements in over 700,000 adults (including prospective studies numbering 550,000 adults in Chennai, Tamil Nadu; 150,000 adults in Mumbai, Maharashtra, and 100,000 adults in Trivandarum, Kerala), and shown that non-medical staff can reliably obtain simple physical measurements. One study (10;11) has surveyed over 100,000 adults in Mumbai and found elevated body mass index [or BMI greater than 25 kg/m2] in approximately 30% of men and women over age 35 and low BMI [<18.5 kg/m2] in some 20% of adults. Thinness was common among illiterate men, and was associated with smoking and chewing tobacco, whereas higher education was associated with raised BMI. The same study has also reported on blood pressure among nearly 89,000 adults (12).

Behavioral pilot studies will focus on HIV-1 risk taking such as sexual partnerships outside marriage. However, such self-reported behavior is greatly misreported (13) and past surveys in India have often had low participation rates. Thus careful pilots will be undertaken prior to a larger study of 5,000 adults in one state. Additional surveys of high-risk populations within SRS areas will also be done to understand spread of HIV-1 and of risk behavior.

Data management

Previous versions of the paper-based VA form were entered (about 40,000 records) in four regional centres in India using Microsoft Access. The written narrative was scanned and retained as a linked image file. A 100% re-check of printouts verified entries. For the most recent round of VA fieldwork (about 110,000 deaths completed in March 2005) and for future entries, central data entry will be performed using scannable, double-sided optical readers. A 100% on-screen re-check of all fields will be done. The written narrative will be retained as an image file for permanent storage and to facilitate re-checks and sub-studies. All VA records are then compiled into a MySQL database for data verification and cleaning.

After scanning, select information (removing personal identifiers) and a complete image of the VA narrative are extracted from the VA database for each record. These fields and an image of the narrative are used to create modified VA reports, entitled “Physician Reports”. Custom-designed, internet-based software permits the electronic distribution and management of physical reports, as well as the remote collection of cause of death

information. Briefly, this system creates Physician Reports from the consolidated VA database, assigns Physician Reports to the appropriate physician (based on language of the narrative and the physician's VA workload), captures the underlying cause of death information, and manages and monitors the administrative tasks related to cause of death coding. This web-based system allows centralized management of the distribution of all VA records, and the secure collection of cause of death data from all parts of India.

Data linkage

In order to better understand the risks for death, each cause-specific death will be linked to its respective baseline record. For the first SRS sample frame, events are linked to the computerized Special Fertility and Mortality Survey (SFMS) of February 1998. For the new SRS sample frame, the linkage will be with the computerized 2004 baseline survey (see below for a list of exposures recorded in each). Additional special surveys will be introduced into the new SRS sample frame for follow up. The new SRS has introduced a unique 9-digit identification number for each individual that can reliably track in-migration, out-migration, vital events and family additions (such as births).

The first SRS sample frame had no such unique identification number. Thus, matching will be based on other variables. Each unique SRS unit is limited to about 150 households. Therefore, matching based on household number, relationship to head of household, gender and name within each of the SRS units can potentially yield a high degree of matching. However, if household numbers are incorrectly assigned in the baseline survey, then manual matching of paper-based forms is likely to be required. A pilot of 389 paper-based SRS death records matched to the SFMS demonstrated that if a paper-based SFMS record could be identified with proper linkage to the death record through the SRS unit number and household number, then matching was successful for 84% of deaths. In-migration (particularly of elderly mothers) and out-migration accounted for an equal proportion of records (about 7% each) that were not successfully matched. Our efforts at data linkage of electronic records have been less successful primarily as a result of variability in the recording of the household number(s) in the paper-based SRS records. We are currently addressing this issue by extracting and linking select variables (such as name) from several SRS forms or schedules, which will enable us to correctly identify SFMS records for most deaths that occur within the SRS sampling frame.

Training of field interviewers and physician coders

Most RGI surveyors are male with at least 12 years of formal (non-medical) education. All RGI surveyors and re-sample teams undergo a standardized, six-day training session in VA methodology. This training is composed of two days of in-class instruction followed by four days of fieldwork, discussion and feedback. VA training includes an introduction to human anatomy and the signs and symptoms of common diseases, mock field interviews with methods to canvas each VA question, hands-on VA writing, and a feedback session to evaluate and improve the training methods. The training aims to improve the surveyors' ability to collect data from the respondents in the open/closed format using symptom checklists and probing questions. The goal is to obtain a complete and logical history of the signs, symptoms and supportive details of each death. The surveyors are trained to seek information from the person with the most details of the illnesses and symptoms prior to death (for all medical causes). For example, for child and neonatal deaths, mothers should be the principle respondent. Each surveyor is required to complete a number of mock VA reports using the techniques and materials provided during the training. Training in VA will be provided as part of the routine activities included within the new SRS sampling frame, and repeated prior to each half-yearly survey. Over 800 RGI surveyors and senior staff have received at least two rounds of training in VA from December 2002 to December 2004.

A network of 15 academic partners from various states work in collaboration with the RGI and the WHO to ensure standardized training, random re-sampling, and to build skills for sustainable mortality measurement.

All VA physicians undergo multiple rounds of training in cause of death assignment. This three-day training covers the importance of VA and how it works, orientation to ICD-10, hands-on exercises in VA, individual work on VA reports in the local language with group discussions on challenging cases, and post-test evaluations/feedback. All physicians have access to web-based training tools, including case-studies, diagnostic guidelines and ICD-10 lists. Senior physician coders review the first 50-100 reports of all new physician coders. The modest honorariums for physicians are paid only on completion of all steps, including reconciliation with another physician. We anticipate training about 150-200 physicians for a cause of death coding panel, depending on

language requirements across the world. As of May 2005, over 95 physicians have been trained.

Research Ethics

SRS enrolment is on a voluntary basis, and its confidentiality and consent procedures are defined as part of the Registration of Births and Deaths Act, 1969. Oral consent was obtained in the first SRS sample frame. The new SRS sample obtains written consent at the baseline. Families are free to withdraw from the study, but the compliance is close to 100%. The study poses no or minimal risks to enrolled subjects. All personal identifiers present in the raw data are anonymized before analysis. The study has been approved by the review boards of the Post-Graduate Institute of Medical Education and Research, the Indian Council of Medical Research, and the Health Ministry's Screening Committee. Specific written consent procedures for additional biological measurements will be added, using international guidelines (14,15).

II. Study organization and study schedule

The study is implemented by a large interdisciplinary team. The RGI-CGHR office in Delhi is responsible for day-to-day management, coordination with states and government offices, and centralized data entry. A Global Coordinating Centre is at the University of Toronto. This Centre provides overall strategic guidance, quality control, and manages the web-based physician coding system. Fifteen Academic Partners in the major Indian states (see list in main report) are responsible for training, coding, and re-sampling in their home states. A national advisory committee comprised of the RGI and project investigators and co-chaired by Professors Vendhan Gajalakshmi and Rajesh Kumar provides input into the project progress.

Independence of the SRS

The SRS is managed by the Office of the Registrar General within the Ministry of Home Affairs of the Government of India. Notably, the Registrar General operates independently of the Ministry of Health and Family Welfare or any disease control programs. This permits *de facto* separation of the producers and potential users of vital registration data.

Role of funding agencies

External funders have no role in study design, data collection, data analysis, data interpretation or writing of publications.

Study schedule

The timetable for the study involves producing two major reports and publications on the baseline characteristics of 2.4 million homes (1.1 million homes in the 1998 SFMS, and 1.3 million homes in the 2004 baseline survey for the New SRS) in 2005. Preliminary cause of death data on about 40,000 deaths at the national level should be available by December 2005, depending on progress in physician coding. The remaining causes of death from the first SRS sample frame will be double coded and analyzed by mid 2006, as will the first 60,000 or so deaths in the new SRS sample frame. Data linkage efforts to link the 1998 SFMS with the deaths from 1998-2003 will be completed by December 2007. Pilot studies of physical and biological measurements surveys within the SRS will begin in fall 2005.

III. Additional validation studies of exposures

Birth order: The percentage of births in the year of the survey that were first, second, third or fourth or higher order births were compared in the bigger states (Assam, Andhra Pradesh, Bihar, Gujarat, Haryana, Himachal Pradesh, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Orissa, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh and West Bengal) between the SFMS and the NFHS-2. Pearson correlations (R^2) across states were 0.94, 0.89, 0.49 and 0.94 for the first, second, third and fourth births, respectively (data not shown).

Indoor air pollution: The use of “dirty fuels” (firewood, dung and crop residue) has been measured as an indicator of indoor air pollution in households in both the SFMS and the 2001 Census. A comparison of the proportion of households that use such dirty fuels for the bigger Indian states reveals a high correlation ($R^2 = 0.8$) between the two surveys (data not shown).

Reference List

(1) Anker M, Black R, Coldham C, Kalter H, Quigley M, et al. (1999) A standard verbal autopsy for investigating causes of death in infants and children. WHO; Report No.: WHO/CDS/CRS/ISR/99.4.

(2) Quigley MA, Chandramohan D, Rodrigues LC (1999) Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. Int J Epidemiol 28:1081-1087.

(3) Kumar R, Sharma AK, Barik S, Kumar V (1989) Maternal mortality inquiry in a rural community of north India. Int J Gynaecol Obstet 1989 29: 313-319.

(4) Gajalakshmi V, Richard P, Santhanakrishnan K, Sivagurunathan B (2002) Verbal autopsy of 48,000 adult deaths attributed to medical causes in Chennai (formerly Madras), India. BMC Public Health. 2:7.

(5) Gajalakshmi V, Peto R (2004) Verbal autopsy of 80,000 adult deaths in Tamil Nadu, South India. BMC Public Health 4: 47.

(6) Kumar R, Thakur J, Rao M, Singh M, Bhatia P (2005) Validity of Verbal Autopsy in Determining Causes of Adult Deaths. Indian Journal of Public Health. In press.

(7) Chandramohan D, Maude GH, Rodrigues LC, Hayes RJ (1988) Verbal autopsies for adult deaths: their development and validation in a multicentre study. Trop Med Int Health 3:436-446.

(8) Kalter HD, Gray RH, Black RE, Gultiano SA (1991) Validation of the diagnosis of childhood morbidity using maternal health interviews. Int J Epidemiol 20:193-198.

(9) WHO (2003) Manual of International Classification of Diseases, Injuries and Causes of Death (Tenth revision). Geneva, Switzerland: World Health Organization.

(10) Gupta PC, Mehta HC (2000) Cohort study of all-cause mortality among tobacco users in Mumbai, India. Bull World Health Organ 78: 877-883.

(11) Shukla HC, Gupta PC, Mehta HC, Hebert JR (2002) Descriptive epidemiology of body mass index of an urban adult population in western India. J Epidemiol Community Health 56: 876-880.

(12) Gupta PC, Gupta R, Pednekar MS (2004) Hypertension prevalence and blood pressure trends in 88,653 subjects in Mumbai, India. J Hum Hypertens 18: 907-910.

(13) Cleland J. Boerma JT, Carael M, Weir SS (2004) Monitoring sexual behaviour in general populations: a synthesis of lessons of the past decade. Sex Transm Infect. 80: Suppl 2 1-7.

(14) Government of India (1998) Revised Guidelines for Exchange of Human Biological Material for Biomedical Research Purposes. Indian Journal of Pharmacology 30: 56-57.

(15) Clayton EW, Steinberg KK, Khoury MJ, Thomson E, Andrews L, Kahn MJ, Kopelman LM, Weiss JO. (1995) Informed consent for genetic research on stored tissue samples. JAMA. 274: 1786-1792.