HYAK **Mortality Monitoring System**

Innovative Sampling and Estimation Methods

Proof of Concept by Simulation

Samuel J. Clark[1,4,5,*], Jon Wakefield[2,3], Tyler McCormick[1,2], and Michelle Ross[2]

[1]Department of Sociology, University of Washington
[2]Department of Statistics, University of Washington
[3]Department of Biostatistics, University of Washington
[4]Institute of Behavioral Science (IBS), University of Colorado at Boulder
[5]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand
[*]Correspondence to: work@samclark.net

September 28, 2012

Working Paper no. 118
Center for Statistics and the Social Sciences
University of Washington

Abstract

Traditionally health statistics are derived from civil and/or vital registration. Civil registration in low- to middle-income countries varies from partial coverage to essentially nothing at all. Consequently the state of the art for public health information in low- to middle-income countries is efforts to combine or triangulate data from different sources to produce a more complete picture across both time and space – what we term *data melding*. Data sources amendable to this approach include sample surveys, sample registration systems, health and demographic surveillance systems, administrative records, census records, health facility records and others. There are very few useful demonstrations of 'data melding', and the two of which we are aware both relate to HIV prevalence.

We propose a new statistical framework for gathering health and population data – Hyak – that leverages the benefits of sampling and longitudinal, prospective surveillance to create a cheap, accurate, sustainable monitoring platform. Hyak has three fundamental components:

- **Data Melding**: a sampling and surveillance component that organizes two data collection systems to work together: (1) data from HDSS with frequent, intense, linked, prospective follow-up and (2) data from sample surveys conducted in large areas surrounding the HDSS sites using informed sampling so as to capture as many events as possible;
- **Cause of Death**: verbal autopsy to characterize the distribution of deaths by cause at the population level; and
- **SES**: measurement of socioeconomic status in order to characterize poverty and wealth.

We conduct a simulation study of the informed sampling component of Hyak based on the Agincourt health and demographic surveillance system site in South Africa. Compared to traditionally cluster sampling, Hyak's informed sampling captures more deaths, and when combined with an estimation model that includes spatial smoothing, produces estimates of both mortality counts and mortality rates that have lower variance and small bias.

We compare the relative cost and precision of Hyak to traditional repeated cluster samples to measure mortality. We find that in as short as two years Hyak is substantially more cost-effective and accurate than current systems.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by the Bill and Melinda Gates Foundation. The authors are grateful to Peter Byass, Basia Zaba, Stephen Tollman, Adrian Raftery, Philip Setel and Osman Sankoh for helpful discussions.

Contents

1 New Directions for Health and Population Statistics in Low- to Middle-Income Countries

1.1 Background

In most of the developed world the traditional source of basic public health information is civil registration. Civil registration is a system that records births and deaths within a government jurisdiction. The purpose is two-fold: (1) to create a legal record for each person, and (2) to provide vital statistics. Optimally a civil register includes everyone in the jurisdiction, provides the basis to ensure their civil rights and creates a steady stream of vital statistics.

The vital statistics obtained from many well-functioning civil registration systems include birth rates by age of mother, mortality rates by sex, age and other characteristics, and causes of death for each death. These basic indicators are the foundation of public health information systems, and when they are taken from a near-full-coverage civil registration system, they relate to the whole population.

Although the idea is inherently simple, implementing full-coverage civil registration is not, and only the world's richest countries are able to maintain ongoing civil registration systems that cover a majority of the population. Civil registration in the rest of the world varies from partial coverage to essentially nothing at all (Mathers et al., 2005). A recent four-article series titled "Who Counts?" in the *Lancet* reviews the current state of civil registration (AbouZahr et al., 2007; Boerma and Stansfield, 2007; Hill et al., 2007; Horton, 2007; Mahapatra et al., 2007; Setel et al., 2007). The authors lament that there has been a half a century of neglect in civil registration in low- to middle-income countries, and critically, that it is not possible to obtain useful vital statistics from those countries (Mahapatra et al., 2007; Setel et al., 2007). The *Lancet* authors argue that in the long term all countries need complete civil registration to ensure the civil rights of each one of their citizens and to provide useful, timely public health information (AbouZahr et al., 2007), and they explore a number of interim options that would allow countries to move from where they are today to full civil registration (Hill et al., 2007). Echoing the *Lancet* special series are addition urgent pleas for better health statistics in low- and middle-income countries (for example: Abouzahr et al., 2010; Bchir et al., 2006; Mathers et al., 2009, 2005; Rudan et al., 2000). These authors clearly identify a need for representative data describing sex-, age-, and cause-specific mortality through time in small enough areas to be meaningful for local governance and health institutions. These critiques are for the most part discussed in the framework of civil registration as the 'primary' source of data.

We agree that in order to ensure civil rights and provide each unique citizen with a legal identity, full-coverage civil registration *is* the long-term goal. Acknowledging that, we believe that the joint discussion of civil registration and vital statistics is not helpful. The conflation of these two topics is an accident of history; the fact that historically civil registration came first and vital registration and vital statistics followed naturally from civil registration.

In recent decades the field of statistics has matured to provide effective and very efficient methods for sampling and using sample data to make inferences about whole populations. *From an information point of view it is no longer necessary to have data describing every vital event and each unique citizen*; in fact, it is wasteful and fiscally irresponsible to argue that full coverage information systems are necessary to produce vital statistics and other public health indicators.

The sample-based approach drives the production of population statistics in many other fields – economics, sociology, political science and many more. Borrowing from these fields public health workers have developed sample-driven approaches to health statistics that partially substitute for vital statistics derived from civil registration. India has conducted a sample registration system (SRS) for several decades (Office of the Registrar General & Census Commissioner, India, 2012) that has produced good basic vital statistics, and more recently Jha and colleagues (2006) have added verbal autopsy (Lopez et al., 2011) to this system to create the Indian Million Death Study (MDS). In a similar vein, USAID's Sample Vital Registration with Verbal Autopsy (SAVVY) is a program that combines sample registration with verbal autopsy and provides general-purpose tools to collect data (MEASURE Evaluation, 2012). USAID's Demographic and Health Surveys (DHS) (Measure DHS, 2012) and UNICEF's Multiple Indicator Cluster Surveys (MICS) (UNICEF - Statistics and Monitoring, 2012) are good examples of traditional household surveys that describe a select subset of indicators for national populations at multiple points in time. There are many more similar sample surveys conducted by smaller organizations and aimed at specific diseases or the evaluation of specific interventions.

Most of these replicate the study designs – specifically the sample designs – developed in other fields to provide cross-sectional snap-shots of the current state of the population with respect to an indicator. With the exception of India's SRS and SAVVY, they lack the ongoing, prospective, longitudinal structure of a traditional vital registration system, and consequently they do not measure changes and trends well. They also often lack the spatial resolution to distinguish differences in indicator values across short distances. Finally, they often miss or undercount rare events because they typically take one measurement and rely on recall to fill in recent history.

The current state of the art for public health information in low- to middle-income countries is efforts to combine or triangulate data from different sources to produce a more complete picture across both time and space. We will refer to this general strategy as *data melding*. The usual sources of data include: non-representative, low-coverage, poor quality vital registration data; roughly once-per-decade census data; snap-shot or repeated snap-shot data from (sometimes nationally representative) household surveys; one-off sample surveys conducted for a variety of specific reasons by a diverse array of organizations; sample registration systems; and finally, a hodgepodge of miscellaneous data sources that may include health and demographic surveillance systems (HDSS), sentinel surveillance systems, administrative records, clinic/hospital records and others.

Alexander Rowe (2009) and Jennifer Bryce (2010) describe a system of 'integrated, continuous surveys' that would produce ongoing, longitudinal monitoring of a variety of outcomes. Data from such a system could be representative with respect to population, time and space and thereby substitute for and improve on traditional vital statistics data. The idea is to systematize the nationally representative household surveys already implemented in a country, conduct them on a regular schedule with a permanent team and institute rigorous quality controls. The innovation is to turn traditional cross sectional surveys into something quasi longitudinal and to ensure a level of consistency and quality. This concept appears to still be in the *idea* stage without any real methodological development or real-world testing. More in the spirit of data melding, Bryce and colleagues (2004) use a variety of data sources to conduct a multi-country evaluation of Integrated Management of Childhood Illness (IMCI) interventions. This evaluation does develop some *ad-hoc* methods for combining and interpreting data from diverse sources.

Victora and colleagues (2011) articulate a similar vision for a national platform for evaluating the effectiveness of public health interventions, specifically those targeting the Millennium Development

Goals (MDG). The authors argue that national coverage with district-level granularity is necessary, and like Rowe and Bryce, that continuous monitoring is required to assess changes and thereby intervention impacts. This article contains significant discussion of general survey methods, sample size considerations and other methodological requirements that would be necessary to evaluate MDG interventions. Again however, there are no methodological details that would allow someone to design and implement a national, prospective survey system of the type described.

Several authors who work at HDSS sites have described an idea for carefully distributing HDSS sites throughout a country in way that could lead to a pseudo representative description of health indicators in the country through time (Ye et al., 2012). Although these authors do not provide details for how this could be done or evidence that it works, the basic idea is supported by work from Byass and his colleagues (2011) who examine the national representativeness of health indicators generated in individual Swedish counties in 1925. Byass and colleagues discover that any of the not-obviously-unusual counties produced indicator values that were broadly representative of the national population – the counties being roughly equivalent to an HDSS site, and Sweden in 1925 being roughly equivalent to low- and middle-income countries today.

Prabhat Jha (2012) summarizes all of this in his description of five ideas for improving mortality monitoring with cause of death. His five ideas include SRS systems with verbal autopsy, improving the representativeness of HDSS (similar to Ye and colleagues (2012)), coordinating, representative retrospective surveys (similar to Rowe and Bryce) and finally using whatever decent-quality civil registration data might be available.

We find only two fully implemented and demonstrated examples of data melding in the public health sphere. Alkema and colleagues (2008; 2007) working with the UNAIDS Reference Group on Estimates, Modelling and Projections develop a Bayesian statistical method that simultaneously estimates the parameters of an epidemiological model that represents the time-evolving dynamics of HIV epidemics *and* calibrates the results of that model to match population-wide estimates of HIV prevalence. The epidemiological model is fit to sentinel surveillance data describing HIV prevalence among pregnant women who attend antenatal clinics, and the population-wide measures of prevalence come from DHS surveys. Interestingly the second example relates to a similar problem. Lanjouw and Ivaschenko at the World Bank (2010) describe a method to meld population-level data from DHS surveys and HIV prevalence data from a sentinel surveillance system. The DHS contains representative information on a variety of items but not HIV prevalence, and the sentinel surveillance system describes the HIV prevalence of a select (non-representative) subgroup, again pregnant women who attend antenatal clinics. Building on ideas in small-area estimation, they develop and demonstrate a method to adjust the sentinel surveillance data and then predict the HIV prevalence of the whole population.

Although these are two specific applications of data melding, *it is this level of conceptual and methodological detail that are necessary in order to meld data from different sources to produce representative, probabilistically meaningful results.* The population, public health and evaluation literatures are full of urgent requests for better data and more useful methods to meld data from different sources to answer questions about *cause and effect* and *change* at national and subnational levels, but there is very little in any of those literatures that actually develops the new concepts and methods that are necessary to deliver the required new capabilities.

1.2 A New Statistical Platform

Taking account of the situation described in the literature and firmly in the spirit of 'data melding', we aim to develop a system that provides high quality, continuously generated, representative vital statistics and other population and health indicators using a system that is cheap and logistically tractable. We are confident that such a system can provide highly useful health information at all important geographical (and other) scales: nation, province, district, and perphaps even subdistrict.

As we argue above, we strongly believe that a *sample-based* approach is both appropriate and sufficient to produce meaningful, useful public health information, and we do not believe it is fiscally responsible to attempt to cover the entire population with a public health information system. That argument must be made on the basis of guaranteeing human rights *alone*.

1.3 Design Criteria

What we want is a *cheap, sustainable*, continuously operated monitoring system that combines the benefits of both sample surveys (representativity, sparse sampling, logistically tractable) and surveillance systems (detailed, linked, longitudinal, prospective with potentially intense monitoring – e.g. of pregnancy outcomes and neonatal deaths) to provide *useful* indicators for large populations over prolonged periods of time, so that we can monitor change and relate changes to possible determinants, including interventions. More specifically, 'useful' in this context means an informative balance of accuracy (bias) and precision (variance) – i.e. minimal but probably not zero bias accompanied by moderate variance. *We want indicators that are close to the truth most of the time*, and we want an ability to study causality properly. Critically, we want the whole system to be cheaper and more sustainable than existing systems, and perhaps offer addition advantages as well.

1.4 Hyak

We propose an integrated data collection and statistical analysis framework for improved population and public health monitoring in areas without comprehensive civil registration and/or vital statistics systems. We call this platform Hyak – a word meaning 'fast' in the Chinook Jargon of the Northwestern United States.

Hyak is conceived as three having three fundamental components:

- **Data Melding**: a sampling and surveillance component that organizes two data collection systems to work together to provide the desired functionality: (1) data from HDSS with frequent, intense, linked, prospective follow-up and (2) data from sample surveys conducted in large areas around the HDSS sites using informed sampling so as to capture as many events as possible.
- **Verbal Autopsy** (Lopez et al., 2011) to estimate the distribution of deaths by cause at the population level, and
- **SES**: measurement of socioeconomic status (SES) at household, and perhaps other levels, in order to characterize poverty and wealth.

Hyak uses relatively small, intensive, longitudinal HDSS sites to understand what types of individuals (or households) are likely to be the most informative if they were to be included in a sample. With this knowledge the areas around the HDSS sites are sampled with preference given to the more informative individuals (households), thus increasing the efficiency of sampling and ensuring that sufficient data are collected to describe rare populations and/or rare events. This fully utilizes the information generated on an on-going basis by the HDSS *and* produces indicator values that are representative of a potentially very large area around the HDSS site(s). Further, the information collected from the sample around the HDSS site can be used to calibrate the more detailed data from the HDSS, effectively allowing the detail in the HDSS data to be extrapolated to the larger population. For an example of how this has been done in the context of antenatal clinic HIV prevalence surveillance and DHS surveys, see Alkema et al. (2008). Another way to do this is to build a hierarchical Bayesian model of the indicator of interest, say mortality, with the HDSS being the first (informative) level and the surrounding areas being at the second level. Thus the surrounding area can borrow information from the HDSS but is not required to match or mirror the HDSS.

In the remainder of this work we focus on the informed sampling component of Hyak. Informed sampling seeks to capture as many events as possible. This is critical for the measurement of mortality, and especially for the measurement of cause-specific mortality fractions (CSMF) at the population level. In order to adequately characterize the epidemiology of a population, it is necessary to measure the CSMF with some precision, and to do this a large number of death events with verbal autopsy are required, especially for rare causes. Informed sampling aims to make the measurement of mortality rates and CSMFs as efficient as possible.

Below we present a detailed example of the informed sampling idea and a pilot study based on information from the Agincourt HDSS site in South Africa (Kahn et al., 2012, 2007). We generate virtual populations based on information from the Agincourt site, and then we simulate applications of traditional one-stage cluster and Hyak sampling designs. We estimate sex-age-specific mortality rates for children ages $0-4$ years (last birthday) and compare and discuss the results.

In the Conclusions Section we describe how verbal autopsy methods can be integrated into the Hyak system and the 'demographic feasibility' of Hyak.

We are thinking about existing data collection methods and these objectives in a unified framework, and we are starting by experimenting with sampling and analysis frameworks that work together to provide the basis for a *measurement system* that is representative, accurate and efficient in terms of information gained per dollar spent (not the same as *cheap* in an absolute sense because estimation of a binary outcome like death is still bound by the fundamental constraints of the binomial model; i.e. relatively large numbers of deaths are needed for useful measurements). To start we are focusing on mortality as our example indicator.

A measurement system like this would be the among the cheapest and most informative ways to monitor the mortality of children affected by interventions that cover large areas and exist for prolonged periods of time, such as the planned Africa Health Markets for Equity (AHME) intervention sponsored by the Gates Foundation and DFID. With this in mind, the pilot project we present below focuses on childhood ages $0-4$.

2 Pilot Study of HYAK Informed-Sampling via Simulation

2.1 Methodological Approach

In this section we describe our approach to sampling and analysis. To be concrete, we suppose that the outcome of interest is *alive* or *dead* for children age $0-4$. There are two novel aspects to our approach:

- **Informed Sampling:** Using existing information from a HDSS site we construct a mortality model based on village-level characteristics. On the basis of this model we subsequently predict the number of outcomes of interest in each village of the study region. We then set sample sizes in each village in proportion to these predictions.
- **Analysis:** We model the sampled deaths as a function of known demographic factors and village-level characteristics, and then we employ spatial smoothing to tune the model to each village and exploit similarities of risk in neighboring villages.

2.1.1 Notation

Given our interest in the binary status *alive* or *dead*, our modeling framework is logistic regression with random effects. Specifically, let $i = 1, ..., I$ represent villages, $j = 1, ..., 4$ index the four levels of sex (F, M) and age categories (< 1 years, $1-4$ years last birthday). Then the quantity of interest is Y_{ij}, the unobserved true number of deaths in village i and in sex/age stratum j. We assume that the populations N_{ij} are known. Also assumed known are village-specific covariates $\boldsymbol{X}_i$. Examples may include the average SES in village i, measures of water quality, and proximity to health care facilities.

The probability of dying in village i and stratum j is denoted p_{ijk}. We stress that we are carrying out a small-area estimation problem so *the target of interest is* Y_{ij} and the probability is just an intermediary which allows us to set up a model. If the full data were observed we would take the probability to be Y_{ij}/N_{ij}.

The survey sample corresponds to choosing n_{ij}, the number of children in stratum j that we sample in village i. Of these y_{ij} are recorded as dying.

2.1.2 Models

In this section we describe models that may be fit to the sampled data. Once we have estimated probabilities $\widehat{p}_{ij}$ we have the estimate:

$$\widehat{Y}_i = y_i + \sum_{j=1}^{4} (N_{ij} - n_{ij}) \times \widehat{p}_{ij}, \tag{1}$$

where y_i is the observed number of deaths and $(N_{ij} - n_{ij})$ is the number of unsampled individuals in village i and stratum j.

I **Naive Model:** The baseline model simply estimates $\widehat{p} = y/n$, i.e. estimates a single probability as the overall empirical risk. The predicted number of deaths in each village is then (1) with $\widehat{p}_{ij} = \widehat{p}$.

II **Age & Sex Model:** This model estimates $\widehat{p}_j = y_j/n_j$, i.e. estimates four probabilities as the empirical risks. The predicted number of deaths in each village is then (1) with $\widehat{p}_{ij} = \widehat{p}_j$.

III **Logistic Regression Covariate Model:** This model fits a model to all villages sampled and estimates stratum effects and in addition the association with village-level covariates $\boldsymbol{x}_i$:

$$\text{logit } p_{ij} = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma_j, \tag{2}$$

where $j = 1, ..., 4$. Hence, we have a model with a separate baseline for each stratum and with the covariates having a common effect across stratum (i.e., no interation between covariates and stratum). Once we have estimates $\widehat{\gamma}_j$ and $\widehat{\boldsymbol{\beta}}$ we can obtain fitted probabilities:

$$\widehat{p}_{ij} = \frac{\exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j)}{1 + \exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j)},$$

which may be used in (1). The model can be fit via maximum likelihood.

IV **Logistic Regression Random Effects Covariate Model:** This model requires all villages to have sampled data and estimates stratum effects and village-level effects as in (2) and in addition introduces random effects to account for unmeasured village-level covariates. Specifically, we assume the model:

$$\text{logit } p_{ij} = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma_j + \epsilon_i + S_i, \tag{3}$$

where $j = 1, ..., 4$. We have two random effects in this model. The *unstructured* error terms $\epsilon_i \sim_{\text{iid}} N(0, \sigma^2_\epsilon)$ are independent and allow for excess-binomial variability. The second set of error terms S_i are spatial random effects that allow the smoothing of rates across space. An obvious way of modeling the spatial random effects is using an intrinsic CAR model (Besag et al., 1991) in which:

$$S_i|S_j, j \in \text{ne}(i) \sim N(\overline{S}_i, \sigma^2_s/n_i),$$

where $\text{ne}(i)$ is the set of neighbors of village i and n_i is the number of such neighbors. To use this model we need a definition of neighbor. Once again we can obtain fitted probabilities:

$$\widehat{p}_{ij} = \frac{\exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j\widehat{\epsilon}_i + \widehat{S}_i)}{1 + \exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j\widehat{\epsilon}_i + \widehat{S}_i)},$$

which may be used in (1). Until recently fitting this model was computationally challenging. However, Rue et al. (2009) have recently described a clever combination of Laplace approximations and numerical integration that can be used to carry out Bayesian inference for this this model. The `inla R` package implements the methods. A Bayesian implementation requires specification of priors for all of the unkown parameters, which for model (3) consist of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2_ϵ and σ^2_s. We choose flat priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and Gamma(a, b) priors for both σ^{-2}_ϵ and σ^{-2}_s.

2.1.3 The Simulation Study Region

We describe the study region that we create for the simulation study, in order to provide a context within which the different sampling strategies can be described. The study region is based on the Agincourt demographic surveillance system (DSS) site in South Africa (Kahn et al., 2007). The

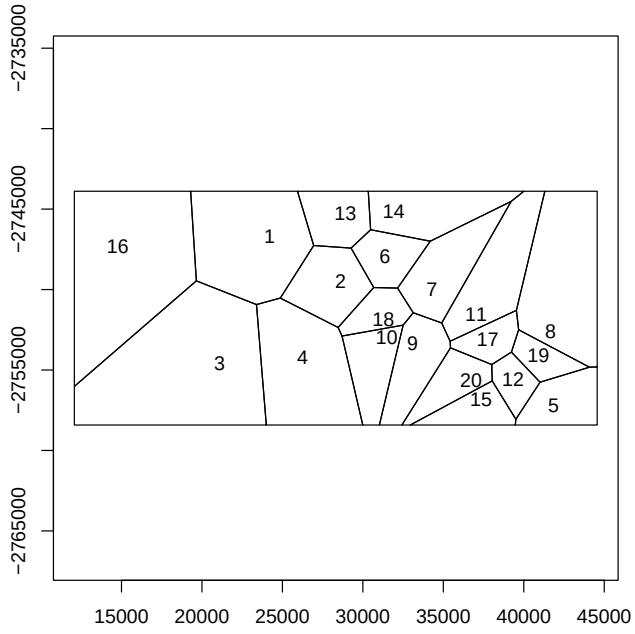

Figure 1: The 20 villages of the Agincourt region with Voronoi tesselations defining neighborhood structure.

total population size is $N = 28,000$ and these individuals reside in one of 20 villages so that there are $N_i = 1,400$ children in each village, of which 700 are girls and 700 boys, with 350 of each age < 1 years, and the other 350 age $1 - 4$ years. For the purpose of this investigation, we treat three of theses villages as Demographic Surveillance System (DSS) sites for which we have extensive and complete information, and the remaining 17 as the 'surrounding area' for which we have only rudimentary information such as would come from a basic census.

We form a Dirichilet tessellation of the village boundaries based on the 20 coordinate pairs that describe the centroids of the villages. This operation forms a set of tiles, each associated with a centroid and is the set of points nearest to that point. This is a standard operation in spatial statstics (e.g. Denison and Holmes, 2001). We can then define neighbors (for the spatial model) as those villages whose tiles share an edge. Figure 1 shows the study region along with village centroids and associated village polygons (as defined by the Voronoi tesselations).

2.1.4 Sampling Strategies

In this section we describe the sampling strategies that we compare. In each strategy the total number of children sampled is 5,200.

- **One-stage Cluster Sampling:** Randomly select 5 villages and sample 1,040 children from each of these villages, 520 girls and 520 boys. This is an example of a one-stage cluster sampling plan, a common design.

- HYAK – **HDSS with Informative Sampling:** We select all children from the 3 HDSS villages and a total of 1,000 from the remaining 17 villages. The numbers of boys and girls from the remaining villages are sampled proportionately to the predicted number of deaths based on the HDSS data. Specifically we fit model (2), and on the basis of that fit (particular the $\boldsymbol{\beta}, \boldsymbol{\gamma}$) we obtain predicted counts of deaths for all villages. Let $\boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star$ be the estimated parameters based on the HDSS data only and $p_{ij}^\star$ be the associated village and stratum-specific parameters. Then the predicted number of deaths in the remaining 17 villages are $\widehat{Y}_{ij} = N_{ij} \times \widehat{p}_{ij}^\star$. We then select sample sizes $n_{ij} \propto \widehat{Y}_{ij}$ so that villages/stratum with more predicted deaths are sampled more heavily. Specifically, if n_{++} is the total sample size we take $n_{ij} = n_{++} \times \widehat{Y}_{ij}/\widehat{Y}_{++}$ where $\widehat{Y}_{++}$ is the total predicted number of deaths (in the non-DSS villages). The observed number of deaths from n_{ij} is y_{ij}.

2.1.5 Measures of Predictive Accuracy

There are different simulation scenarios that may be envisaged. Given the N total children, equally broken into the four stratum, we can set risks p_{ij} for each village/stratum and then simulate counts Y_{ij}. We can then subsample from these counts, under each of the four designs and repeat $s = 1, .., S$ times. Taking this one step further we might wish to average over the possible realizations of counts Y_{ij}.

The estimated number of deaths in survey villages in simulation s is

$$\widehat{Y}_{ij}^{(s)} = y_{ij}^{(s)} + (N_{ij} - n_{ij}) \times \widehat{p}_{ij}^{(s)}$$

where the $\widehat{p}_{ij}^{(s)}$ are obtained from one of the models we described in Section 2.1.2.

We now define the accuracy measures that we use. An obvious measure of accuracy is the mean squared error (MSE) associated with the predicted number of deaths. The MSE, averaged over villages and strata is

$$\begin{aligned} \text{MSE} &= \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{20}\sum_{j=1}^{4}(Y_{ij} - \widehat{Y}_{ij}^{(s)})^2 && (4) \\ &= \sum_{i=1}^{20}\sum_{j=1}^{4}(\overline{\widehat{Y}}_{ij} - Y_{ij})^2 + \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{20}\sum_{j=1}^{4}(\widehat{Y}_{ij}^{(s)} - \overline{\widehat{Y}}_{ij})^2 && (5) \\ &= \sum_{i=1}^{20}\sum_{j=1}^{4}\text{Bias}(\widehat{Y}_{ij})^2 + \sum_{i=1}^{20}\sum_{j=1}^{4}\text{Var}(\widehat{Y}_{ij}). && (6) \end{aligned}$$

where

$$\overline{\widehat{Y}}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\widehat{Y}_{ij}^{(s)}$$

is the average of the predicted counts over simulations in village i and stratum j. The decomposition in terms of **bias** and **variance** is useful since it makes apparent the trade-off involved in modeling.

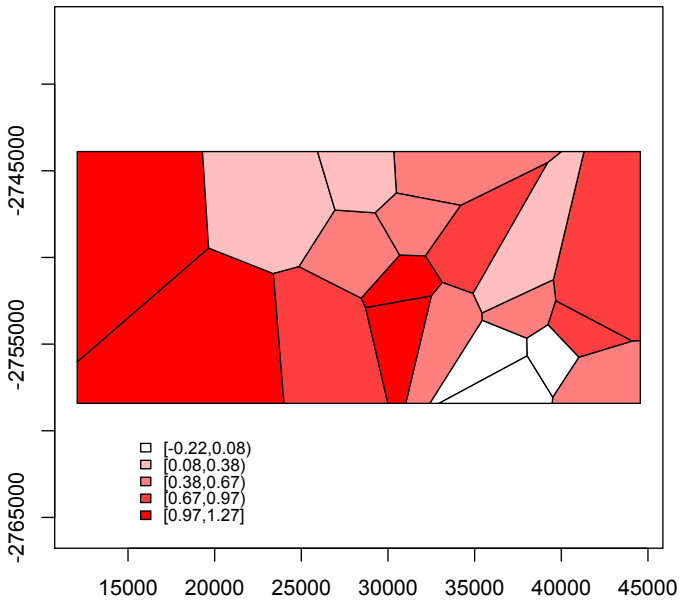

Figure 2: The simulated spatial random effects for the Agincourt region.

2.2 Simulation

We assume there are two village-level covariates so that the length of the $\boldsymbol{\beta}$ vector is 2. Both of the village-level covariates x_{i1} and x_{i2} are generated independently from uniform distributions on 0 to 1. Based loosely on the real values from the Agincourt HDSS in South Africa, the parameter values we use in the simulation are:

- The risk of death in young girls is $\text{expit}(\gamma_1) = 0.050$.[1]
- The risk of death in young boys is $\text{expit}(\gamma_2) = 0.117$.
- The risk of death in older girls is $\text{expit}(\gamma_3) = 0.032$.
- The risk of death in older girls is $\text{expit}(\gamma_4) = 0.071$.
- The first village-level covariate has $\exp(\beta_1) = \exp(-1.1) = 0.333$ so that a unit increase in x_1 leads to the odds of death dropping by a third.
- The second village-level covariate has $\exp(\beta_2) = \exp(0.7) = 2.01$ so that a unit increase in x_2 leads to the odds of death doubling.
- We set $\sigma^2_\epsilon = 0.22$ to determine the level of unstructured variability. This leads to a 95% range for the residual unstructured odds being $\exp(\pm 1.96 \times \sqrt{0.22}) = [0.40, 2.51]$.
- We set $\sigma^2_s = 0.48$ to determine the level of unstructured variability. This operation requires

[1] expit() is the inverse of logit()

some care because the ICAR model does not define a proper probabiity distribution. The ICAR variance is not directly interpretable and so instead Figure 2 shows a simulated set of S_i, $i = 1, ..., 20$ values, with darker values indicating higher risk. The spatial dependence is apparent, with this realization producing high risk to the West of the region and low risk in the East.

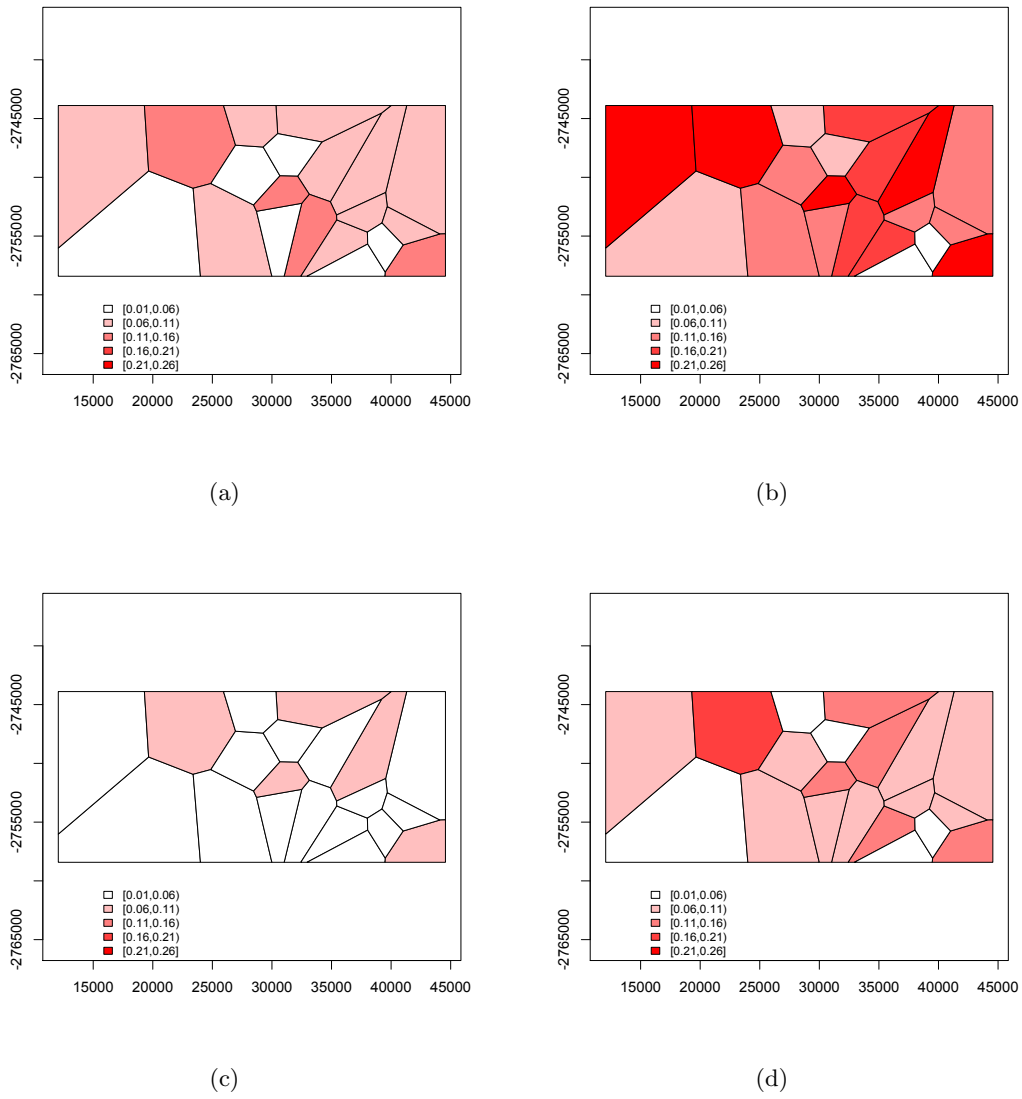

Figure 3: The predicted probabilities for the Agincourt region: (a) young girls, (b) young boys, (c) older girls, (d) older boys.

Combining all of the elements of the model, we generate deaths Y_{ij} for village i and stratum j by randomly drawing from a Binomial distribution with probabilities given by (3). This yields the predicted probabilities for all 20 villages and for each of the four stratum displayed in Figure 3.

The HDSS villages are selected by taking the villages with both large x_1 and large x_2, small x_1 and small x_2, followed by a randomly sampled third village.

A $Ga(5, 1)$ prior is used for both the spatial and non-spatial random effects in the spatial models.

2.3 Results

Table 1 summarizes the results of the simulation study. The number of sampled deaths and bias, variance and MSE from (6) are displayed for each combination of sampling strategy and analytical model. Table 2 contains the differences for each metric comparing the cluster sample and HYAK – raw differences and proportional differences with the cluster sample as the base for comparison.

Overall regardless of analytical model, it is clear that the HYAK sampling strategy captures more deaths and is more accurate – the MSE is smaller in general. Further examining the components of the MSE, what emerges is: (i) HYAK yields far, far lower variance, and (ii) pays for this by sacrificing some bias, but not much. The overall comparison between the sampling strategies very clearly favors HYAK. This partly reflects the careful choice of HDSS villages so that they contain substantial variation in terms of village-level covariates.

Comparing the analytical models also produces and encouraging result. All of the models perform better overall (smaller MSEs) when using the HYAK sampling strategy, and within HYAK, model IV *logistic regression random effects covariate model* or *Covariates & Space* outperforms the others. This suggests that accounting for unmeasured factors and taking advantage of the spatial structure of mortality risk is significantly worthwhile.

Table 1: Deaths, Bias, Variance, MSE for Cluster Sample & HYAK

Sampling	Model	Deaths	Bias	Variance	MSE
Cluster	I. Naïve	473	40	353	1,918
	II. Age & Sex	473	39	349	1,888
	III. Covariates	473	39	439	1,981
	IV. Covariates & Space	473	-na-	-na-	-na-
Hyak	I. Naïve	549	42	86	1,855
	II. Age & Sex	549	42	85	1,823
	III. Covariates	549	42	85	1,820
	IV. Covariates & Space	549	41	94	1,780

Results from 100 simulations; 2,578 deaths. 'Cluster' is shorthand for *One-stage Cluster Sample*; 'HYAK' for *HDSS with Informative Sampling*; 'Covariates' for *Logistic Regression Covariate Model* and 'Covariates & Space' for *Logistic Regression Random Effects Covariate Model*. It is not possible to fit the spatial model (IV) to the one-stage sampling plan since there are data from 5 villages only.

Table 2: Comparisons: HYAK vs. Cluster Sample

Comparison	Model	Deaths	Bias	Variance	MSE
Differences	I. Naïve	76	3	-268	-63
	II. Age & Sex	76	3	-264	-65
	III. Covariates	76	2	-354	-162
	IV. Covariates & Space	76	-na-	-na-	-na-
Proportional Differences	I. Naïve	16%	6%	-76%	-3%
	II. Age & Sex	16%	6%	-76%	-3%
	III. Covariates	16%	6%	-81%	-8%
	IV. Covariates & Space	16%	-na-	-na-	-na-

Differences are: HYAK − Cluster
Proportional Differences are: $\frac{\text{HYAK}-\text{Cluster}}{\text{Cluster}}$

The trade-off between bias and variance is clearly revealed by a closer look at the distributions of the estimated probability of dying produced by each model. Figure 4 displays these distributions for models I, III & IV – *Naïve, Covariates* and *Covariates & Space* under the HYAK sampling strategy. The *Naïve* model estimates are very spread out, always include the truth and have some bias; estimates from the *Covariates* model have very little spread, almost always miss the truth and have clear bias; and finally, estimates from the *Covariates & Space* model have spread that is intermediate between the other two models, distributions that always include the truth, and about the same bias. Clearly the *Covariates & Space* model displays the balance we are seeking: manageable bias, small spread, and importantly, distributions that always include the truth. This combination of sampling strategy and analytical approach provides our key objective: an indicator that is close to (and around) the truth most of the time.

3 Costing HYAK: Relative Precision & Expenditures

In this section we present the estimated financial commitment required to establish and maintain HYAK. We also compare the relative cost and precision of HYAK to other mortality monitoring systems currently in use. We find that in as short as two years, HYAK is substantially more cost-effective and accurate than current systems.

3.1 Estimated Commitment for HYAK

Data collection using HYAK involves two parts: a central surveillance unit (similar to a HDSS site) and statistically informed sampling outside the central site. Developing the initial infrastructure for HYAK is similar to establishing a lightweight HDSS site, which we approximate at between $250,000 and $500,000, depending on local circumstances. This cost covers hiring personnel (a relatively modest permanent staff and part-time enumerators), equipment, and permanent physical space. Some permanent locations will likely be within existing facilities (in a hospital or clinic, for example) which may reduce up-front costs. The infrastructure developed through the initial costs produces significant savings in subsequent per-respondent costs. We estimate that the cost to visit

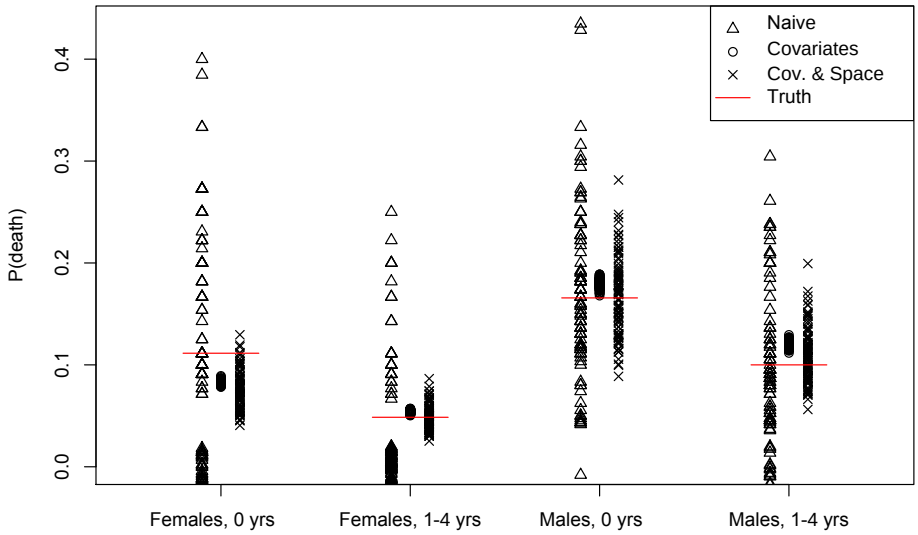

Figure 4: Example of Bias / Variance Trade-Off.

a respondent in the central site will be about one-third of the cost to survey a respondent on a typical platform, such as the DHS (Tollman and Ezeh, 2012).

Data collection at centralized HDSS-like sites would then be supplemented by surveys outside the central site. These surveys would require a basic census of the entire region to establish a sampling frame. Respondents would then be selected with probabilities determined using the statistically informed sampling techniques described in Section 2.1.4. Statistically informed sampling increases the propensity of selecting individuals in populations of interest, which increases efficiency and reduces uncertainty.

3.2 Cost Comparison

Using approximate cost figures from the Agincourt HDSS site in South Africa and the APHRC HDSS in Kenya (Tollman and Ezeh, 2012), we estimate the financial commitment to establish and maintain a HYAK system compared to other data collection platforms. We estimate that data collection within the central site would cost about ∼$7.50 per individual in each round of data collection and year. Respondents selected using statistically informed sampling would cost about ∼$40 per individual (Ezeh, 2012) (the APHRC HDSS in Nairobi, Kenya has conducted survey samples similar to what is necessary for HYAK for a cost of ∼$40 per subject). The initial lightweight census to determine a sampling frame would be ∼$20 per individual.

In computing cost estimates, we use population sizes from the simulations in Section 2.1.3 (N=28,000 individuals in 20 villages). We keep the total sample size for both HYAK and a DHS-like method the same. We assume that the DHS-like method does an initial census to establish a sampling-frame at a cost of $20 per population member and samples each respondent at a cost of $40 per individual. For comparison, we use the simulation design which uses 260 children in each village,

which approximates the geographically clustered sampling design used by the DHS. For Hyak, we take all individuals in three villages as the central surveillance site with a start-up cost of $325,000 and survey cost of $7.50. Additionally, we do statistically informed sampling for 1000 individuals in different villages. The cost for these individuals is the same as that of the DHS-like method. We assume data collection proceeds for five years with two rounds per year.

Figure 5 displays the cumulative cost for Hyak and a DHS-like system for five years. The statistically informed sampling used by Hyak produces lower MSE than a cluster-based sampling design, as demonstrated in Section 2.3. Hyak is also less expensive after the initial investment period to establish the central surveillance site. In Figure 5 the cumulative investment is approximately equal in year two. Establishing more central surveillance sites would, of course, increase the required initial investment and, thus, increase the time needed for the costs of the two approaches to be comparable. Using the figures we have provided, however, we anticipate that the costs would likely still be comparable within a reasonable time period (using three central sites in our set-up would increase the time to only seven years, for example).

3.3 Overall Cost Assessment

Overall these results demonstrate that a system with permanent infrastructure such as Hyak is financially more well-suited for sustainable, long-term monitoring than a cluster-based sampling design. Further, statistically informed sampling produces greater accuracy than a cluster-based sampling scheme by increasing the propensity for selecting members of the population of interest.

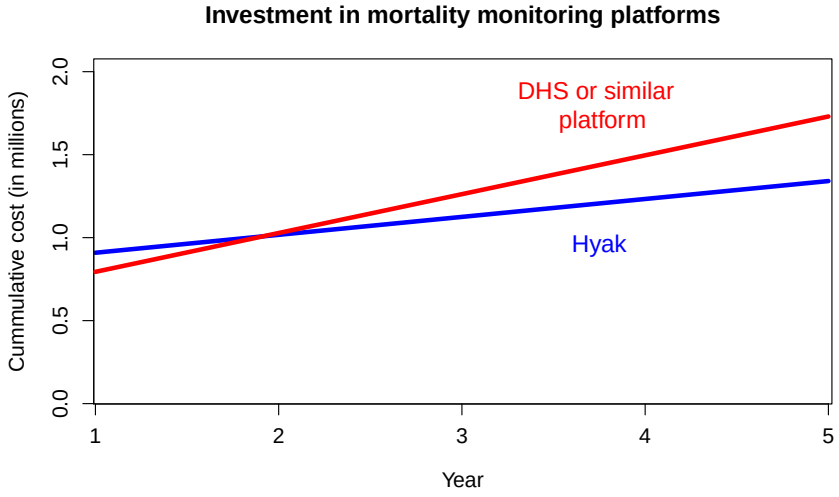

Figure 5: Cumulative cost for Hyak and a DHS-like system for five years.

4 Discussion

4.1 Key Conclusions

The key conclusion of this pilot study is that the statistical sampling and analysis ideas supporting the Hyak monitoring system are sound: a combination of highly informative data such as are produced by a HDSS site can be used to judiciously inform sampling of a large surrounding area to yield estimated counts of deaths that are far more useful than those produced by a traditional cluster sample design. Further, Hyak combined with an analytical model that includes unstructured random effects and spatial smoothing produces the most accurate and well-behaved estimates. The improvements are dramatic and clearly justify additional work on these ideas.

Another crucial idea underlying Hyak is the notion that very detailed information generated by a HDSS site can be extrapolated to the much larger surrounding population by calibrating that information with carefully chosen and much less detailed data from the surrounding population. This idea has already been demonstrated convincingly by Alkema et al. [2008] and is currently being applied by UNAIDS to produce global estimates of HIV prevalence.

A key advantage of Hyak sampling strategy is that it *captures significantly more deaths*. Verbal autopsy methods (Lopez et al., 2011) can be applied to all or a fraction of these deaths to assign causes (immediate, contributing, etc.). This cause of death information can then be used to construct distributions of deaths by cause – CSMFs – which illuminate the epidemiological regime affecting the population, and if this is monitored through time, how the epidemiology of the population is changing. Critically, this provides a means of measuring the impact of interventions on specific causes of death and the distribution of deaths over time. The increased number of deaths captured with informed sampling increases the accuracy and precision of measurements of CSMFs.

A final benefit of the Hyak system is that it provides two types of infrastructure: the HDSS and the sample survey. In addition to providing information with which to sample, the HDSS provides a platform on which a wide variety of longitudinal studies can be undertaken – linked observational studies; randomized, controlled trials, all kinds of combinations of these, etc. Moreover, the permanent HDSS infrastructure also provides a training platform that can support a wide variety of health and behavioral science training, mentoring and apprenticing/interning and experience for young scientists of health professionals. Having the sample survey infrastructure provides a means of quickly validating/calibrating studies conducted by the HDSS and provides another learning dimension for the educational and training activities that the system can support.

A potential limitation of any mortality monitoring system is 'demographic feasibility', that is the ability to capture enough deaths in a given population to measure levels and/or changes in mortality, potentially by cause, through time. Death is a binomial process defined by a probability of dying, and as such, is governed by the relatively simple characteristics of the binomial model. That model specifies in simple terms the number of deaths necessary to estimate the probability of dying within a given margin of error with a given level of confidence. No amount of sophistication will release us from that basic set of facts. The Hyak system addresses this challenge by providing a means through which to choose the best possible sample given what we know about the population, and this in turn maximizes our ability to capture deaths. The fundamentals of the binomial model require that one must observe relatively large numbers of deaths to measure mortality precisely and especially to measure changes in mortality with both precision and confidence. So in light

of those inescapable realities, the HYAK system produces the most information per dollar spent, because it captures more deaths per dollar spent.

Finally and perhaps most importantly, the HYAK monitoring system is cheaper to run over a period of years compared to traditional cluster sample-based survey methods. Combined with the fact that HYAK also produces more useful information, this makes HYAK highly cost effective – *more bang for less buck.*

4.2 Future Work

Bringing the HYAK idea to fruition will require a lot of additional work, likely:

- More simulation studies that incrementally approximate a wider variety (high and low mortality, various covariates and spatial relationships, etc.) of real situations more faithfully. These can be informed by historical data sets at the national level (e.g. Sweden) and HDSS sites in low- to middle-income countries.
- A small 'real-world' pilot study that aims to implement all of the logistics necessary for HYAK and work out the kinks so that we know how to run a HYAK system efficiently and effectively.
- A full scale pilot implementation that establishes a HYAK system for a population that is large enough to be meaningful, i.e. ∼1 million people. It will likely be advantageous to do this in partnership with an existing HDSS site in order to save the cost of setting up a new HDSS site.

References

AbouZahr, C., J. Cleland, F. Coullare, S. B. Macfarlane, F. C. Notzon, P. Setel, S. Szreter, R. N. Anderson, A. a. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, T. Evans, X. C. Figueroa, C. K. George, L. Gollogly, R. Gonzalez, D. R. Grzebien, K. Hill, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, P. Jha, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, A. D. Lopez, D. M. Fat, M. Merialdi, L. Mikkelsen, J. K. Nien, C. Rao, K. Rao, O. Sankoh, K. Shibuya, N. Soleman, S. Stout, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, and S. Zhang (2007, November). The way forward. *Lancet 370*(9601), 1791–9.

Abouzahr, C., L. Gollogly, and G. Stevens (2010, February). Better data needed: everyone agrees, but no one wants to pay. *Lancet 375*(9715), 619–21.

Alkema, L., A. Raftery, and T. Brown (2008). Bayesian melding for estimating uncertainty in national hiv prevalence estimates. *Sexually transmitted infections 84*(Suppl 1), i11–i16.

Alkema, L., A. Raftery, and S. Clark (2007). Probabilistic projections of hiv prevalence using bayesian melding. *The Annals of Applied Statistics 1*(1), 229–248.

Bchir, A., Z. Bhutta, F. Binka, R. Black, D. Bradshaw, G. Garnett, K. Hayashi, P. Jha, R. Peto, C. Sawyer, B. Schwartländer, N. Walker, M. Wolfson, D. Yach, and B. Zaba (2006, January). Better health statistics are possible. *Lancet 367*(9506), 190–3.

Besag, J., J. York, and A. Molliè (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics 43*, 1–59.

Boerma, J. T. and S. K. Stansfield (2007). Health Statistics 1 Health statistics now : are we making the right investments ? *Tuberculosis 369*(9563), 779–786.

Bryce, J. and R. Steketee (2010, February). Continuous surveys and quality management in low-income countries: a good idea. *The American journal of tropical medicine and hygiene 82*(2), 360; author reply 361–2.

Bryce, J., C. G. Victora, J.-P. Habicht, J. P. Vaughan, and R. E. Black (2004, March). The multi-country evaluation of the integrated management of childhood illness strategy: lessons for the evaluation of public health interventions. *American journal of public health 94*(3), 406–15.

Byass, P., O. Sankoh, S. M. Tollman, U. Högberg, and S. Wall (2011, January). Lessons from history for designing and validating epidemiological surveillance in uncounted populations. *PloS one 6*(8), e22897.

Denison, D. and C. Holmes (2001). Bayesian partitioning for estimating disease risk. *Biometrics 57*(1), 143–149.

Ezeh, A. (2012). Personal Communication with Director of APHRC HDSS, Kenya.

Hill, K., A. D. Lopez, K. Shibuya, P. Jha, C. AbouZahr, R. N. Anderson, A. a. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, J. Cleland, F. Coullare, T. Evans, X. Carrasco Figueroa, C. K. George, L. Gollogly, R. Gonzalez, D. R. Grzebien, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, D. M. Fat, S. Macfarlane, P. Mahapatra, M. Merialdi, L. Mikkelsen, J. K. Nien, F. C. Notzon, C. Rao, K. Rao, O. Sankoh, P. W. Setel, N. Soleman, S. Stout, S. Szreter, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, S. Zhang, and M. Zhou (2007, November). Interim measures for meeting needs for health sector data: births, deaths, and causes of death. *Lancet 370*(9600), 1726–35.

Horton, R. (2007, November). Counting for health. *Lancet 370*(9598), 1526.

Jha, P. (2012). Counting the dead is one of the world's best investments to reduce premature mortality. *Hypothesis 10*(1).

Jha, P., V. Gajalakshmi, P. C. Gupta, R. Kumar, P. Mony, N. Dhingra, and R. Peto (2006, February). Prospective study of one million deaths in India: rationale, design, and validation results. *PLoS medicine 3*(2), e18.

Kahn, K., M. Collinson, F. Gómez-Olivé, O. Mokoena, R. Twine, P. Mee, S. Afolabi, B. Clark, C. Kabudula, A. Khosa, et al. (2012). Profile: Agincourt health and socio-demographic surveillance system. *International Journal of Epidemiology 41*(4), 988–1001.

Kahn, K., S. Tollman, M. Collinson, S. Clark, R. Twine, B. Clark, M. Shabangu, F. Gómez-Olivé, O. Mokoena, and M. Garenne (2007). Research into health, population and social transitions in rural south africa: Data and methods of the agincourt health and demographic surveillance system1. *Scandinavian Journal of Public Health 35*(69 suppl), 8–20.

Lanjouw, P. and O. Ivaschenko (2010). A new approach to producing geographic profiles of hiv prevalence. Technical Report 5207, World Bank - Policy Research Working Paper Series.

Lopez, A. D., R. Lozano, C. J. Murray, and K. Shibuya (2011). Verbal autopsy: innovations, applications, opportunities - improving cause of death measurement (article collection). *Population Health Metrics 9*.

Mahapatra, P., K. Shibuya, A. D. Lopez, F. Coullare, F. C. Notzon, C. Rao, and S. Szreter (2007, November). Civil registration systems and vital statistics: successes and missed opportunities. *The Lancet 370*(9599), 1653–1663.

Mathers, C. D., T. Boerma, and D. Ma Fat (2009, January). Global and regional causes of death. *British medical bulletin 92*, 7–32.

Mathers, C. D., D. M. Fat, M. Inoue, C. Rao, and A. D. Lopez (2005, March). Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the World Health Organization 83*(3), 171–7.

Measure DHS (2012). Demographic and Health Surveys. `http://www.measuredhs.com`.

MEASURE Evaluation (2012). SAVVY: Sample Vital Registration with Verbal Autopsy. `http://www.cpc.unc.edu/measure/tools/monitoring-evaluation-systems/savvy`.

Office of the Registrar General & Census Commissioner, India (2012). India's Sample Registration System. `http://censusindia.gov.in/Vital_Statistics/SRS/Sample_Registration_System.aspx`.

Rowe, A. K. (2009, June). Potential of integrated continuous surveys and quality management to support monitoring, evaluation, and the scale-up of health interventions in developing countries. *The American journal of tropical medicine and hygiene 80*(6), 971–9.

Rudan, I., J. Lawn, S. Cousens, A. K. Rowe, C. Boschi-Pinto, L. Tomasković, W. Mendoza, C. F. Lanata, A. Roca-Feltrer, I. Carneiro, J. a. Schellenberg, O. Polasek, M. Weber, J. Bryce, S. S. Morris, R. E. Black, and H. Campbell (2000). Gaps in policy-relevant information on burden of disease in children: a systematic review. *Lancet 365*(9476), 2031–40.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology) 71*(2), 319–392.

Setel, P. W., S. B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, and C. AbouZahr (2007, November). A scandal of invisibility: making everyone count by counting everyone. *Lancet 370*(9598), 1569–77.

Tollman, S. and A. Ezeh (2012). Personal Communication with Directors of Agincourt HDSS, South Africa and APHRC HDSS, Kenya.

UNICEF - Statistics and Monitoring (2012). Multiple Indicator Cluster Surveys (MICS). `http://www.unicef.org/statistics/index_24302.html`.

Victora, C. G., R. E. Black, J. T. Boerma, and J. Bryce (2011, January). Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. *Lancet 377*(9759), 85–95.

Ye, Y., M. Wamukoya, A. Ezeh, J. Emina, and O. Sankoh (2012, September). Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa? *BMC public health 12*(1), 741.