global health, epidemiology and genomics

STATISTICAL METHODS
ORIGINAL RESEARCH ARTICLE

HYAK mortality monitoring system: innovative sampling and estimation methods – proof of concept by simulation

S. J. Clark[1,2,3,4]*, J. Wakefield[5,6], T. McCormick[5,7] and M. Ross[8]

[1] *Department of Sociology, The Ohio State University, Columbus, Ohio, USA*
[2] *MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), Faculty of Health Sciences, University of the Witwatersrand, School of Public Health, Johannesburg, South Africa*
[3] *INDEPTH Network, Accra, Ghana*
[4] *ALPHA Network, London, UK*
[5] *Department of Statistics, University of Washington Seattle, Washington, USA*
[6] *Department of Biostatistics, University of Washington, Seattle, Washington, USA*
[7] *Department of Sociology, University of Washington, Seattle, Washington, USA*
[8] *Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA*

Global Health, Epidemiology and Genomics (2018), **3**, e3, page 1 of 14. doi:10.1017/gheg.2017.15

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 – *data amalgamation*. Data sources amenable to this approach include sample surveys, sample registration systems, health and demographic surveillance systems, administrative records, census records, health facility records and others. 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 amalgamation*: A sampling and surveillance component that organizes two or more 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 Health and Demographic Surveillance System (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
- *Socioeconomic status (SES)*: Measurement of SES in order to characterize poverty and wealth.

We conduct a simulation study of the informed sampling component of HYAK based on the Agincourt HDSS site in South Africa. Compared with traditional 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.

Received 30 September 2017; Revised 13 October 2017; Accepted 16 October 2017

Key words: Mortality measurement, data amalgamation, sampling, HDSS, sample survey.

* Address for correspondence: S. J. Clark, Department of Sociology, The Ohio State University, Ohio, USA.
(Email: work@samclark.net)

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

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 and vital statistics. Civil registration is a system that records births and deaths within a government jurisdiction. The purpose is twofold: (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 [1].

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 [2]. A four-article series titled 'Who Counts?' in the *Lancet* in 2007 reviews the current state of civil registration [3–8]. This was followed eight years later with another four-article series presenting a similar but slightly more hopeful picture [9–12]. 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 [7,8,11].

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 [3,9], and they explore a number of interim options that would allow countries to move from where they are today to full civil registration [5]. Echoing the *Lancet* special series are additional urgent pleas for better health statistics in low- and middle-income countries [for example: 2, 13–16]. The WHO and its partners and supporters have actively supported improvements in civil registration and vital statistics (CRVS) over the recent past [17–19]. These workers 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.

Recently, various United Nations (UN) agencies, including the office of the Secretary General, have articulated strong, specific support for rapid improvement in the evidence base for the Sustainable Development Goals (SDG) [20] – the international target framework that follows from the Millennium Development Goals (MDGs) [e.g., 21–23]. The appropriately named *Data Revolution* [24] is the flagship program organized by the UN to address the systematic lack of data to measure progress toward the SDG targets.

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 propose decoupling the discussion of civil registration from vital statistics. In particular, we can obtain accurate and representative vital statistics measurements by making inferences from carefully adjusted samples.

The sample-based approach drives the production of population statistics in many other fields, including economics, sociology, and political science. 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 [25] that has produced good basic vital statistics, and more recently Jha *et al.* [26] have added verbal autopsy [27] to this system to create the Indian MDS (Million Death Study). 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 [28]. USAID's Demographic and Health Surveys (DHS) [29] and UNICEF's Multiple Indicator Cluster Surveys (MICS) [30] 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.

These approaches generally utilize sampling designs developed to provide cross-sectional snapshots 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. 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 involves efforts to combine or triangulate data from multiple sources to produce a more complete picture across both time and space. The usual sources of data include: non-representative, low-coverage, poor quality vital registration data; roughly once-per-decade census data; snapshot or repeated snapshot 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

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

that may include health and demographic surveillance systems (HDSS), sentinel surveillance systems, administrative records, clinic/hospital records, and others.

Combining data from different sources with multiple sampling schemes present a myriad of statistical challenges. We use *data pooling* as a broad term that describes methods that adjust for bias due to differences in representativeness across data from different sources. The global burden of disease study by the Institute for Health Metrics and Evaluation [31] is a highly visible example of data pooling. As another example, [32] pool survey data to produce fine geographical scale *Plasmodium falciparum* malaria endemicity. *Data amalgamation* also uses data from multiple sources but is differentiated by active engagement in the data collection process. Data amalgamation uses proactive (e.g. Hyak) or adaptive mechanisms that actively adjust the data collection process to optimize a set of metrics – minimize bias, minimize variance, minimize cost, etc. A recent study of malaria prevalence by [33] is an example of *data amalgamation* in which survey locations are adaptively chosen to minimize the variance of a target; see [34] for statistical details. In the survey sampling literature, adaptive cluster sampling has a relatively long history [35], and has been used extensively in surveys of rare animal and plant species; we are not aware of any applications in the context considered here. In short, we use data pooling for situations where researchers combine several datasets not necessarily collected to measure the indicator of interest, whereas data amalgamation is an intentional strategy that incorporates multiple heterogeneous data sources into the design process.

In the spirit of earlier work on master sampling frames, the United Nations Statistical Division [36–38] describe a system of 'integrated, continuous surveys' that would produce ongoing, longitudinal monitoring of a variety of outcomes – a proactive engagement with the data collection process in keeping with our definition of *amalgamation*. 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 amalgamation, Bryce *et al.* [39] 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 *et al.* [40] articulate a similar vision for a national platform for evaluating the effectiveness of public health interventions, specifically those targeting the 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. That 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 [41]. 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 *et al.* [42] 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.

Jha [43] 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 *et al.* [41]), 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 amalgamation in the public health sphere. Before the DHS surveys routinely collected HIV biomarkers that could be used to estimate population HIV prevalence at the national level, several groups developed methods that drew on multiple sources of data to generate reasonable national-level estimates of HIV prevalence. Alkema *et al.* [44,45] 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 [46] describe a method to amalgamate population-level data from DHS surveys and HIV prevalence data from a sentinel surveillance system. The DHS contains representative

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

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 amalgamation, *it is this level of conceptual and methodological detail that are necessary in order to amalgamate 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 amalgamate data from different sources to answer questions about *cause and effect* and *change* at the national and subnational levels, but there is very little in any of those literatures that actually develop the new concepts and methods that are necessary to deliver the required new capabilities. Chipeta et al. [34] describe an adaptive design whose aim is to estimate disease prevalence.

1.2. A new statistical platform

Taking account of the situation described in the literature and firmly in the spirit of 'data amalgamation', 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 perhaps 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 additional 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 having three fundamental components:

- *Data amalgamation*: 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* [27] to estimate the distribution of deaths by cause at the population level, and
- *Socioeconomic status (SES)*: Measurement of 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.* [44]. 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

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

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[1] in South Africa [47,48]. The Agincourt HDSS is situated in the rural northeast of South Africa and covers an area of 420 km^2 comprising a subdistrict of 27 villages. The site monitors roughly 90 000 people in 16 000 households. The villages and households are dispersed widely across this area, and there is a functional road network linking them all. The epidemiology of the site is typical for South Africa with generally low mortality except for the effect of HIV at very young and middle ages, and in terms of wealth/poverty, the population is typical of a middle-income country [e.g. 49–54]. The Agincourt HDSS is the canonical HDSS, not extreme along any dimension, and generally representative of what an HDSS site is.

We generate virtual populations based on information from the Agincourt site, and then we simulate applications of traditional two-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 analytical 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).

A measurement system like this would be 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. 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 $i = 1, \ldots, I$ represent villages within the study region and $j = 1, \ldots, 4$ index strata, which we take as the four levels of sex (F, M) and age (Young: [0, 1) years, old: [1, 5) years). Households within areas will be represented by $k = 1, \ldots, K_i$, for $i = 1, \ldots, I$. 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 in all villages. Also assumed known are village-specific covariates $\boldsymbol{X}_i$ (for example, the average SES in village i, a measure of water quality, or proximity to health care facilities).

The probability of dying in village i and stratum j is denoted by p_{ij}, which is the hypothetical proportion of children dying in a hypothetical infinite population in area i and strata j. 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 the observed frequency $\tilde{p}_{ij} = Y_{ij}/N_{ij}$. The

[1] From [47]: The Agincourt health and socio-demographic surveillance system (HDSS), located in rural northeast South Africa close to the Mozambique border, was established in 1992 to support district health systems development led by the post-apartheid ministry of health. At baseline in 1992, 57 600 people were recorded in 8900 households in 20 villages; by 2006, the population had increased to 70 000 people in 11 700 households. This increase is partly due to Mozambican in-migrants overlooked in the baseline survey and to a new settlement established as part of the post-apartheid governments Reconstruction and Development Program. In 2007, the study area was extended to include the catchment area of a new privately supported community health centre established to provide HIV treatment before public sector roll-out of HAART. By mid-2011, the population under surveillance comprised 90 000 people residing in 16 000 households in 27 villages. Households are self-defined as people who eat from the same pot of food. Given sustained high levels of temporary labor migration in southern Africa, we included temporary migrants residing for <6 months/year who retain close ties with their rural homes in the HDSS. There have been 17 census and vital event update rounds conducted strictly annually since 2000. Participation is virtually complete, with only two households refusing to participate in 2011.

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

survey design problem 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.

In the next section, we describe models that will be used to analyze the data; once we have estimated probabilities from a generic model, $\widehat{p}_{ij}$, we use the estimator:

$$\widehat{Y}_{ij} = y_{ij} + (N_{ij} - n_{ij}) \times \widehat{p}_{ij}, \tag{1}$$

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

2.1.2. Models

In this section, we describe models that may be fit to the sampled data.

I *Naïve model:* This baseline model simply estimates $\widehat{p} = y/n$, i.e. a single probability is applied to the unsampled individuals in each village. The predicted number of deaths in each village is then (1) with $\widehat{p}_{ij} = \widehat{p}$.

II *Strata model:* This model estimates $\widehat{p}_j = y_j/n_j$, so that estimates of four stratum-specific probabilities are calculated. The predicted number of deaths in each village is then (1) with $\widehat{p}_{ij} = \widehat{p}_j$.

III *Covariate model:* This approach fits a model to data from all villages where sampling was carried out and estimates stratum effects along with the association between risk and village-level covariates x_i. We assume a logistic form,

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

where $j = 1, \ldots, 4$. Hence, we have a model with a separate baseline for each stratum and with the covariates having a common effect across stratum and village, so there is no interaction between covariates and stratum, and covariates and area. We use the maximum-likelihood estimates $\widehat{\gamma}_j$ and $\widehat{\boldsymbol{\beta}}$ to obtain fitted probabilities:

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

which may be used in (1).

IV *Spatial covariate model:* This approach requires sufficient villages to have sampled data, so that spatial random effects can be estimated. Specifically, we assume a Bayesian implementation of the model:

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

where $j = 1, \ldots, 4$. We have three random effects in this model. The *unstructured* village- and household-level error terms $\epsilon_i \sim_{iid} N(0, \sigma_\epsilon^2)$ and $h_k \sim_{iid} N(0, \sigma_h^2)$, respectively, are independent and allow for excess-binomial variability. The household-level random effects also allow for dependence within households. The S_i error terms are village-level spatial random effects that allow the smoothing of rates across space. There are many different forms that these random effects could take. A model-based *geostatistical approach* [55] would assume the collection $[S_1, \ldots, S_n]$ arise from a multivariate normal distribution, with covariances a function of the distance between villages. We go a different route and use an intrinsic conditional auto-regressive (ICAR) model [56] in which:

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

where ne(i) is the set of neighbors of village i and n_i is the number of such neighbors. This model assumes that the prior distribution for the spatial effect in area i, given its neighbors, is centered on the mean of the neighbors, with a variance that depends on the number of neighbors (with more neighbors reducing the prior variance). We describe our 'shared boundary' neighborhood scheme in the next section. We use the posterior means $\widehat{\boldsymbol{\beta}}, \widehat{\gamma}_j, \widehat{\epsilon}_i$, and $\widehat{S}_i$ to 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); we do not include the household random effects as these are not relevant to predicting an area-level summary, but rather account for within-household clustering. Until relatively recently, fitting this model was computationally challenging within the context of a simulation study (which requires repeated fitting). However, Rue *et al.* [57] have described a clever combination of Laplace approximations and numerical integration that can be used to carry out Bayesian inference for this model – the integrated nested Laplace approximation (INLA). The INLA R package implements the INLA method. A Bayesian implementation requires specification of priors for all of the unknown parameters, which for model (3) consist of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ_ϵ^2, σ_s^2, and σ_h^2. We choose flat priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and Gamma(a, b) priors for σ_ϵ^{-2}, σ_s^{-2}, and σ_h^{-2}.

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 HDSS site in South Africa [47,48]. We assume N individuals reside in one of 20 villages and that there are between 1400 and 14 000 children in each village, $N_i \sim \text{Unif}(1400, 14\,000)$. In addition for each village, we assume half the children are boys and half are girls, with 20% in the age range 0–1 years and 80% in the age range 1–5 years. Within each village, we assume that households contain between one and five children and follow the distribution

- P(household with one child) = 75/470 = 0.16
- P(household with two children) = 100/470 = 0.21
- P(household with three children) = 125/470 = 0.27

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

- *P*(household with four children) = 100/470 = 0.21
- *P*(household with five children) = 70/470 = 0.15.

We sample a single population of *N* children and then take *S* = 100 repeated draws from this population under the four sampling schemes described below. Beginning with the denominators N_{ij}, we sample the observed deaths y_{ij} using a binomial with probabilities given by (2).

We sample a second population of *N* children and treat this population as a historical cohort. It is from this population that we treat three of these villages as HDSS sites for which we have extensive and complete information.

We form a Voronoi 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 statistics [e.g. 58]. 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), along with edges showing the neighborhood structure.

2.1.4. Sampling strategies

In this section, we describe the sampling strategies that we compare. In each strategy, we consider four different sample sizes, *n*, for the total number of children sampled: 1300, 2600, 3900, and 5200.

- *Two-stage cluster sampling:* This is the design most commonly used by the DHS, MICS and similar household surveys. Randomly select five villages and randomly sample (*n*/5)/3 households within each of the villages (since each household contains, on average, three children). Additional households will be sampled as needed until at least *n*/5 children are obtained from each village. This is an example of a two-stage cluster sampling plan, a common design.
- *Stratified sampling:* Randomly sample *n*/20 children from each of the 20 villages. This strategy lies between the cluster sampling and informed sampling designs.

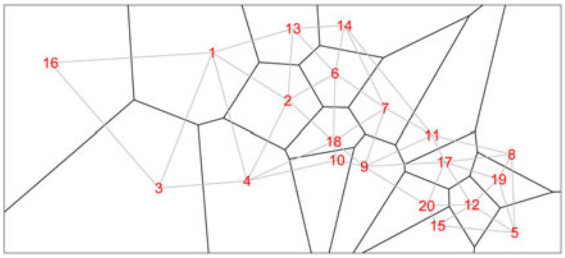

Fig. 1. The 20 villages of the Agincourt region with Voronoi tesselations defining neighborhood structure. Gray lines indicate neighboring villages.

- *Hyak – HDSS with informative sampling:* The number of children sampled from each village is proportional to the predicted number of deaths based on the HDSS data. In particular, we select all children from the three HDSS villages in the historical cohort and we fit model (2). On the basis of the estimated $\boldsymbol{\beta}$, γ, we obtain predicted counts of deaths for all villages, using the village-level covariates $\boldsymbol{x}_i$, $i = 1, \ldots, I$. Let $\boldsymbol{\beta}^*$, γ^* be the estimated parameters based on the historic HDSS data only and p^*_{ij} be the associated village and stratum-specific probabilities. We estimate p_i^* via

$$p_i^* = \sum_{j=1}^{J} \frac{N_{ij}}{N_i} p_{ij}^*.$$

Then, the predicted number of deaths are $\widetilde{Y}_i = N_i \times p_i^*$. We then select sample sizes as the (rounded versions of) $n_i \propto \widetilde{Y}_i$ so that villages with more predicted deaths are sampled more heavily. Specifically, we take $n_i = n \times \widetilde{Y}_i / \widetilde{Y}_+$ where $\widetilde{Y}_+$ is the total predicted number of deaths. The observed number of deaths from n_i is y_i.
- *Optimum allocation:* As in the Hyak sampling design, we obtain the village-level estimates of the probability of death, p_i^*, based on the historic HDSS data only. We then select sample sizes as the (rounded versions of)

$$n_i = n \times \frac{N_i\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}{\sum_{i'} N_{i'}\sqrt{\widehat{p}_{i'}(1-\widehat{p}_{i'})}}.$$

Details are provided in the appendix.

2.1.5. Measures of predictive accuracy

Given *N* total children, broken into the four stratum, we can set risks p_{ij} (details of which appear in the section 'Simulation') for each village/stratum and then simulate counts Y_{ij}. We take this set of $\{Y_{ij} : i = 1, \ldots, 20; j = 1, \ldots, 4\}$ as fixed, and then subsample from these counts, under each of the four designs and repeat $s = 1, \ldots, S$ times.

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 the section 'Models'.

To estimate the frequentist properties of the simulation procedure, we summarize the results by examining various summary measures. An obvious measure of accuracy is the mean-squared error (MSE) associated with the predicted number of deaths. The MSE of an estimator of the number of deaths in area *i* and strata *j*, $\widehat{Y}_{ij}$ averaged over villages and strata is

$$\text{MSE}(\widehat{Y}_{ij}) = E\left[\left(y_{ij} - \widehat{Y}_{ij}\right)^2\right],$$

where y_{ij} is the true number of deaths (which recall, is fixed), and the expectation is over all possible samples that can be taken (for whichever design we are considering). This MSE is estimated based on S simulations:

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

where Y_{ij} is the true number of deaths in village i and stratum j and

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

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.

2.2. Simulation

We assume that there are two village-level covariates so that the length of the β vector is 2. Both of the village-level covariates x_{i1} and x_{i2} are generated independently from uniform distributions on 0 to 1, $i = 1, \ldots, 20$. 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 expit(γ_1) = 0.050.
- The risk of death in young boys is expit(γ_2) = 0.117.
- The risk of death in older girls is expit(γ_3) = 0.032.
- The risk of death in older boys is expit(γ_4) = 0.077.
- The first village-level covariate has $\exp(\beta_1) = \exp(-2.2) = 0.111$, so that a unit increase in x_1 leads to the odds of death dropping by one-ninth.
- The second village-level covariate has $\exp(\beta_2) = \exp(1.4) = 4.05$, so that a unit increase in x_2 leads to the odds of death quadrupling.
- We set $\sigma_\epsilon^2 = 0.22$ to determine the level of unstructured variability at the village level. This leads to a 95% range for the residual unstructured village-level odds being $\exp(\pm 1.96 \times \sqrt{0.22}) = [0.40, 2.51]$.
- We set $\sigma_s^2 = 0.48$ to determine the level of structured variability at the village level. This operation requires some care because the ICAR model does not define a proper probability distribution. The ICAR variance is not interpretable as a marginal variance (and so is not comparable to the other random effects variances, σ_ϵ^2 and σ_h^2) and so instead Fig. 2 shows a simulated set of S_i, $i = 1, \ldots, 20$ 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.

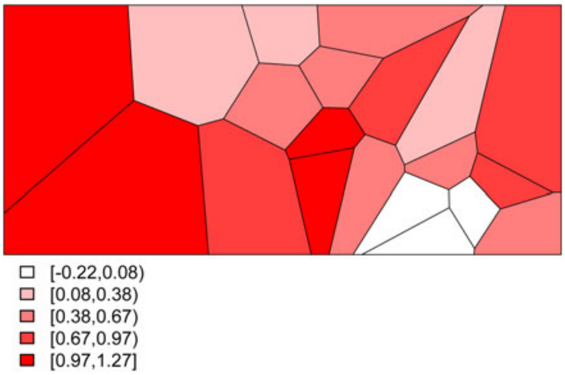

Fig. 2. The simulated spatial random effects for the Agincourt region.

- We set $\sigma_h^2 = 0.08$ to determine the level of unstructured variability at the household level. This leads to a 95% range for the residual unstructured household-level odds being $\exp(\pm 1.96 \times \sqrt{0.08}) = [0.57, 1.74]$.

For the strata and covariates models, the covariate relationship is estimated from the villages that produced data, and then model (2) is used to obtain fitted probabilities that are applied to the unsampled villages, using the population and covariate information that is assumed known for each village.

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 Fig. 3. The historic cohort is generated in the same fashion. Details of the village-level characteristics for both cohorts are provided in Appendix A.2.

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 Gamma(5, 1) prior is used for the spatial and non-spatial random effects in the spatial models (model IV).

2.3. Results

Table 1 summarizes the results of the simulation study for n = 5200. Results for the smaller sample sizes are shown in Tables A3–A5 in Appendix A.3 The number of average sampled deaths and bias, variance and MSE from (4) are displayed for each combination of sampling strategy and analytical model.

Overall, the Hyak sampling strategy captures more deaths and is generally more accurate. Across sampling schemes and sample sizes, Hyak generally has the smallest MSEs. Further examination of the components of the MSE reveals that: (i) Hyak yields smaller bias, and (ii) pays for this by sacrificing some variance. The overall comparison between the

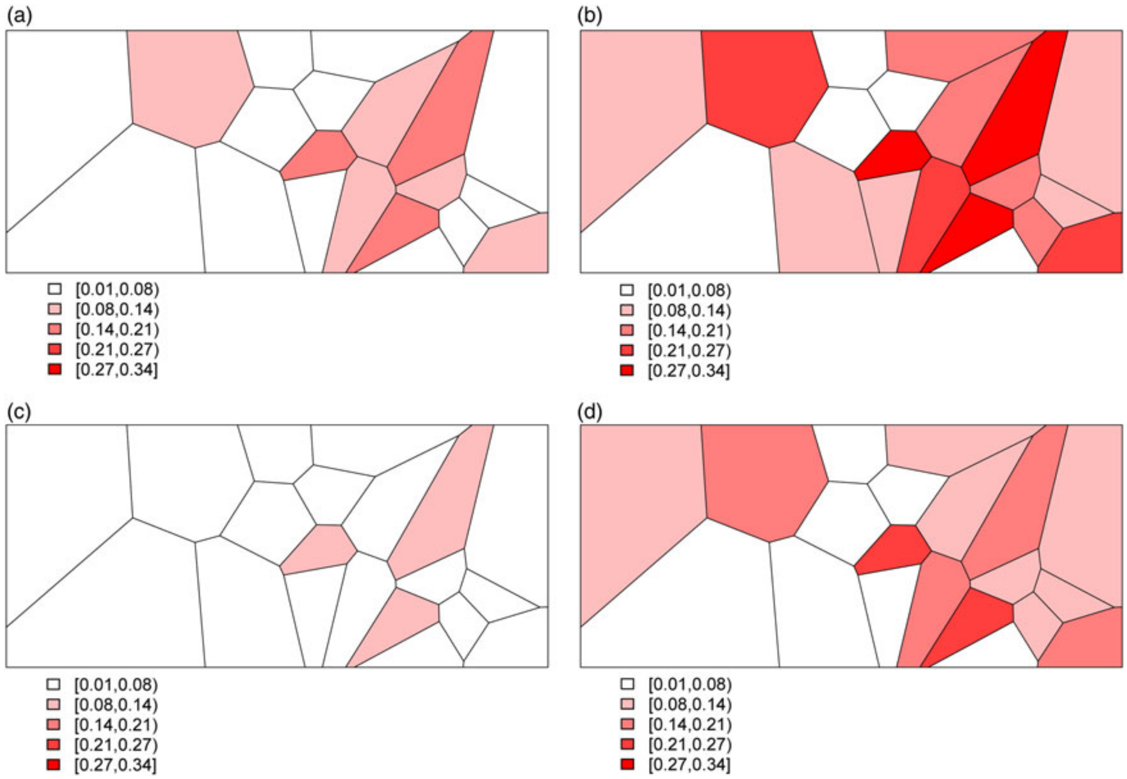

Fig. 3. The predicted probabilities of dying for the Agincourt region: (a) young girls, (b) young boys, (c) older girls, (d) older boys.

Table 1. *Deaths, bias, variance, MSE for cluster sampling, stratified sampling, Hyak, and optimum sampling for* n = *5200*

Design	Model	Deaths	Bias	Variance $(\times 10^3)$	MSE $(\times 10^3)$
Cluster	I. Naïve	459	1067	174	1312
	II. Strata	459	874	188	951
	III. Strata/Covariates	459	651	386	810
	IV. Strata/Covariates/Space	459	–	–	–
Stratified	I. Naïve	460	1058	5	1124
	II. Strata	460	866	15	765
	III. Strata/Covariates	460	651	16	439
	IV. Strata/Covariates/Space	460	183	80	113
Hyak	I. Naïve	538	1162	7	1357
	II. Strata	538	969	18	956
	III. Strata/Covariates	538	635	16	419
	IV. Strata/Covariates/Space	538	182	66	100
Optimum	I. Naïve	477	1072	5	1154
	II. Strata	477	880	18	792
	III. Strata/Covariates	477	632	17	416
	IV. Strata/Covariates/Space	477	167	74	102

Results from $S = 100$ simulations. There were 11 299 deaths in the simulated population from which samples were taken. ‘Cluster’ is shorthand for Two-stage Cluster Sample; ‘Hyak’ for HDSS with Informative Sampling; ‘Strata/Covariates’ for Logistic Regression Covariate Model and ‘Strata/Covariates/Space’ for Logistic Regression Random Effects Covariate Model. It is not possible to fit the spatial model (IV) to the two-stage cluster sampling scheme since there are data from five villages only.

sampling strategies 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 an encouraging result. Within each sampling strategy, the Logistic Regression Random Effects Covariate model (model IV) performs best overall (smaller MSEs). Within HYAK, this outperforms the others. Similar patterns are observed across all sample sizes. This suggests that accounting for unmeasured factors and taking advantage of the spatial structure of mortality risk is significantly worthwhile.

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, and IV – *Naïve, Covariates and Covariates & Space* under the HYAK sampling strategy for $n = 5200$, while Figs A.1–A.3 in Appendix A.3 display these same distributions for $n = 3900$, $n = 2600$, and $n = 1300$, respectively. The *Naïve* model estimates are very condensed, always miss the truth and have clear bias; estimates from the *Covariates* model also have very little spread, almost always miss the truth and have some bias; and finally, estimates from the *Covariates & Space* model have large spread; however, the distributions nearly always include the truth, and have much less bias. Clearly the *Covariates & Space* model displays the balance we are seeking: small bias and manageable spread, and importantly, distributions that 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.

Figure 5 displays the average village- and strata-specific estimates for the (unobserved) population counts of death plotted against the true values across each of the four models under the HYAK sampling scheme for $n = 5200$, while Figs A.4–A.6 in Appendix A.3 display the same for the smaller sample sizes. (See Figs A.7–A.18 in Appendix A.3 for the remaining sampling schemes for all sample sizes.) In general, the average estimates from the spatial model tend to follow the $y = x$ line quite closely, indicating we are estimating the true number of deaths in each village quite well. Estimates tend to be closer under the HYAK sampling strategy and for larger sample sizes, thus confirming (visually) our previous results.

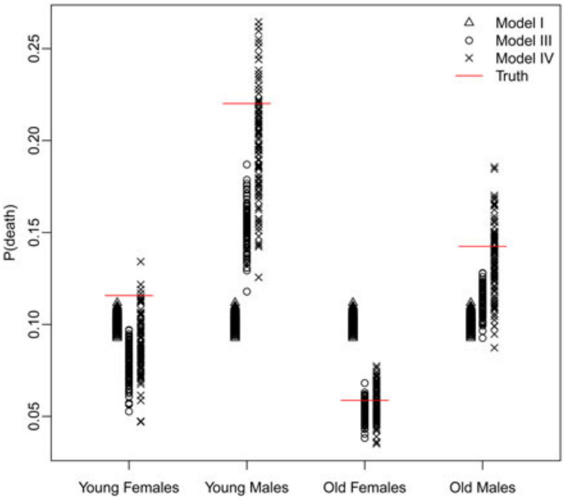

Fig. 4. The distributions of the estimated probability of dying from models I, III, and IV under the HYAK sampling strategy for $n = 5200$.

3. Discussion

3.1. 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 those 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 an 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.* [44] and is currently being applied by UNAIDS to produce global estimates of HIV prevalence. This relies on the assumption that the population monitored by the HDSS is similar enough to the population surrounding the HDSS that the relationships between covariates and the outcomes of interest are the same or very similar. The degree to which this is true will vary among specific settings. In particular, when HDSS sites also serve as research and intervention testing sites, it is possible that there will be *Hawthorne Effect* issues – i.e. the intensively studied HDSS population will be different from the surrounding population that has not participated in studies and trials. This may affect the key covariate–outcome relationships that drive HYAK. This is something that must be studied, initially with a real-world pilot study of HYAK, and then in an ongoing way by occasionally verifying these relationships through an over-sample of the surrounding population, or through small add-on studies conducted whenever a census is done in the surrounding areas to update the sampling frame. Although this is a concern, it is unlikely to make HYAK infeasible or invalidate HYAK results. An explicit goal of a pilot study will be to characterize the uncertainty created by

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

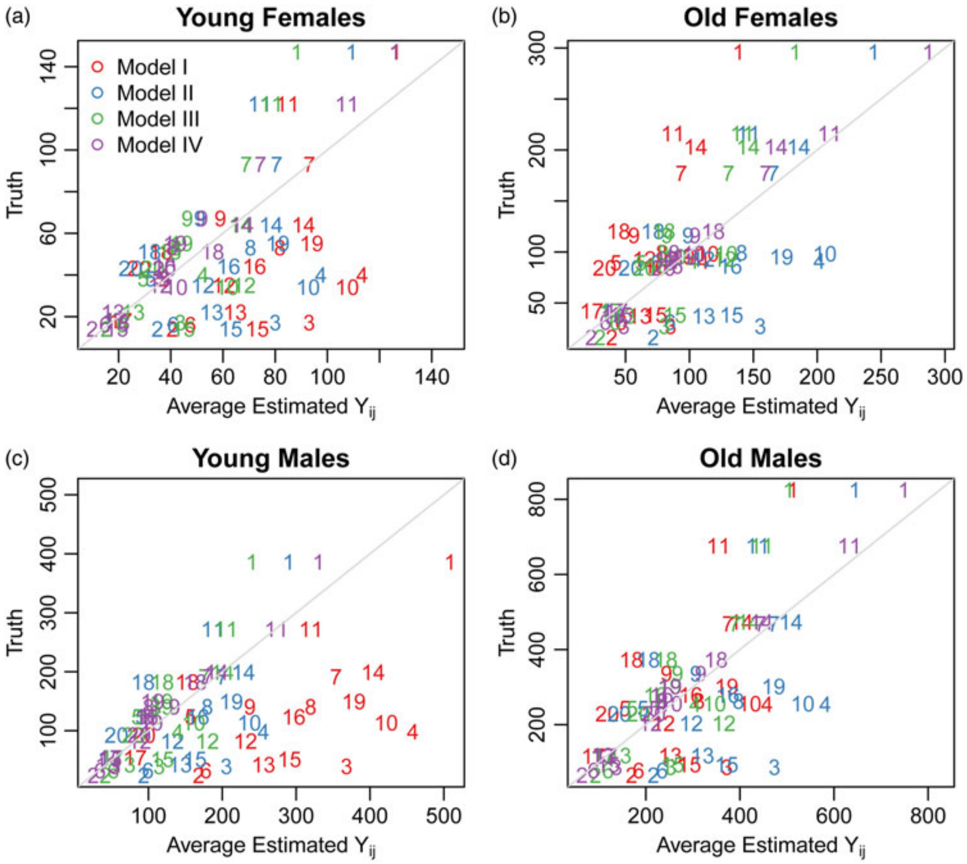

Fig. 5. The average village- and strata-specific estimates for the (unobserved) population counts of death plotted against the true values across each of the four models under the Hyak sampling scheme for $n = 5200$. Plotting symbols indicate village numbers, and colors indicate model number with key in the upper-left plot. The spatial model IV (purple) symbols are in general closest to the $y = x$ line of equality.

possible Hawthorne Effect issues and build them into Hyak estimates.

A key advantage of Hyak sampling strategy is that it *captures significantly more deaths*. Verbal autopsy methods [27] 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 or 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 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

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press

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 with the 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*.

Importantly, there are implementation considerations that must be addressed before Hyak can be used at provincial or national scale to provide population-representative estimates. These will need to be resolved through additional theoretical work, simulation, and ultimately through a pilot study that conducts Hyak on a large population dispersed over a large physical space. Among many, these critical questions need to be answered:

- How big do HDSS sites need to be to provide enough information for effective informative sampling?
- How many HDSS sites are necessary for effective informed sampling with respect to key demographic and epidemiological indicators?
- How should HDSS sites be dispersed geographically?
- How well does Hyak work to provide disaggregated (fine-grained) estimates of key indicators by sex, age, wealth/ poverty, space, time, etc?
- How much does the sampling frame affect Hyak results, and what cheap, feasible solutions are there to obtaining frequently updated sampling frames?
- A detailed costing and cost comparison needs to be done comparing the costs of the HDSS site; the additional census, sampling, and interviewing needed for Hyak; and a traditional household multi-stage cluster sample survey (such as DHS) conducted in the same area.
- How the method can be scaled up to a larger geographical area. We envisage that only a subset of villages will be sampled, and then a geostatistical model [59] can be used for spatial prediction to unobserved villages (a critical question is the number of villages needed to train the spatial model). Another important issue is to deal with the potential problem of preferential sampling [60] in which sampling locations are selected based on the expected size of the response. In order to inform sampling historical data (e.g., DHS surveys) may be used to model to create a predictive surface, upon which sampling may be based. Investigating this idea will be the subject of a future paper.
- When confronted with the potential to locate new HDSS sites so that they are maximally useful in a Hyak-type setup in the future, the most important question is *where exactly to locate the HDSS sites*. A partial answer motivated by the design of Hyak is that an HDSS site should be located in each (mostly) homogeneous region – so that it is reasonable to assume that the relationship between the Hyak selection variable (SES in the work we present here) and the outcomes of interest is the same throughout the Hyak area, critically including the HDSS. What needs further theoretical and practical investigation is how similar the HDSS and surrounding areas need to be, and how Hyak estimates break down as the HDSS and surrounding area become less similar.

Finally, the rapid improvement of CRVS systems is an important priority for many developing countries and funders who support the production of population and health data. The Hyak idea has the potential to play an important role in the transition to fully functioning CRVS in many such countries. Using the same combination of informed sampling and spatiotemporal smoothing, it may be possible to leverage different sources of data to produce good estimates of the vital rates that one expects from a high-quality vital statistics system. A variety of administrative records could stand in for the HDSS site, and sampled vital registration systems could play the role of the household survey. Both types of data collection system already exist and function well in some developing country settings. Thinking through, simulating, and pilot testing this type of integration is likely the highest pay-off next step in developing Hyak.

Acknowledgments

Preparation of this manuscript was supported by the Bill and Melinda Gates Foundation and grants K01 HD057246 and K01 HD078452 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). JW was supported by R01-CA095994. The authors are grateful to Peter Byass, Basia Zaba, Stephen Tollman, Adrian Raftery, Philip Setel, Osman Sankoh, and two very constructive anonymous reviewers for helpful discussions or other inputs.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/gheg.2017.15

References

1 **United Nations. Statistical Division**. *Principles and Recommendations for a Vital Statistics System*, 3rd edn. New York: United Nations Department of Economic and Social Affairs, 2014.

2 **Mathers CD, *et al.***. 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* 2005; **83**: 171–177.

3 **AbouZahr C**, ***et al*.**. The way forward. *Lancet* 2007; **370**: 1791–1799.

4 **Boerma JT, Stansfield SK.** Health Statistics 1 Health statistics now : are we making the right investments ? *Tuberculosis* 2007; **369**: 779–786.

5 **Hill K**, ***et al*.**. Interim measures for meeting needs for health sector data: births, deaths, and causes of death. *Lancet* 2007; **370**: 1726–1735.

6 **Horton R.** Counting for health. *Lancet* 2007; **370**: 1526.

7 **Mahapatra P**, ***et al*.**. Civil registration systems and vital statistics: successes and missed opportunities. *Lancet* 2007; **370**: 1653–1663.

8 **Setel PW**, ***et al*.**. A scandal of invisibility: making everyone count by counting everyone. *Lancet* 2007; **370**: 1569–1577.

9 **AbouZahr C**, ***et al*.**. Towards universal civil registration and vital statistics systems: the time is now. *Lancet* 2015a; **386**: 1407–1418.

10 **AbouZahr C**, ***et al*.**. Civil registration and vital statistics: progress in the data revolution for counting and accountability. *Lancet* 2015b; **386**: 1373–1385.

11 **Mikkelsen L**, ***et al*.**. A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *Lancet* 2015; **386**: 1395–1406.

12 **Phillips DE**, ***et al*.**. Are well functioning civil registration and vital statistics systems associated with better health outcomes? *Lancet* 2015; **386**: 1386–1394.

13 **Abouzahr C, Gollogly L, Stevens G.** Better data needed: everyone agrees, but no one wants to pay. *The Lancet* 2010; **375**: 619–621.

14 **Bchir A**, ***et al*.**. Better health statistics are possible. *Lancet* 2006; **367**: 190–193.

15 **Mathers CD, Boerma T, Ma Fat D.** Global and regional causes of death. *British Medical Bulletin* 2009; **92**: 7–32.

16 **Rudan I**, ***et al*.**. Gaps in policy-relevant information on burden of disease in children: a systematic review. *Lancet* 2000; **365**: 2031–2040.

17 **World Health Organization**. Strengthening civil registration and vital statistics for births, deaths and causes of death: resource kit. Technical Report, World Health Organization, 2013.

18 **World Health Organization**. Strengthening civil registration and vital statistics systems through innovative approaches in the health sector: guiding principles and good practices. Technical Report, World Health Organization, 2013.

19 **World Health Organization**. Improving mortality statistics through civil registration and vital statistics systems: strategies for country and partner support. Technical Report, World Health Organization, 2014.

20 **United Nations**. Sustainable Development Goals, 2014. (http://sustainabledevelopment.un.org/owg.html).

21 **Commission on Population and Development**. Strengthening the demographic evidence base for the post-2015 development agenda: report of the secretary-general. Technical report, United Nations, 2016.

22 **Data Revolution Group: The UN Secretary General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development**. A world that counts: Mobilising the data revolution for sustainable development. Technical Report, United Nations, 2014.

23 **United Nations**. Resolution 2016/1: Strengthening the demographic evidence base for the 2030 agenda for sustainable development, 2016. (http://undocs.org/E/CN.9/2016/1).

24 **United Nations**. Data Revolution for Sustainable Development, 2014a. (http://www.un.org/apps/news/story.asp?NewsID=48594#.VEVQpoctuvJ).

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

26 **Jha P**, ***et al***. Prospective study of one million deaths in India: rationale, design, and validation results. *PLoS Medicine* 2006; **3**: e18.

27 **Lopez AD**, ***et al*.**. Verbal autopsy: innovations, applications, opportunities – improving cause of death measurement (article collection). *Population Health Metrics* 2011; **9** (https://www.biomedcentral.com/collections/verbal-autopsy).

28 **MEASURE Evaluation**. SAVVY: Sample Vital Registration with Verbal Autopsy, 2012. (http://www.cpc.unc.edu/measure/tools/monitoring-evaluation-systems/sav vy).

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

30 **UNICEF – Statistics and Monitoring**. Multiple Indicator Cluster Surveys (MICS), 2012. (http://www.unicef.org/statistics/index_24302.html).

31 **Naghavi M**, ***et al*.**. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet* 2015; **385**: 117–171.

32 **Gething PW**, ***et al*.**. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal* 2011; **10**: 378.

33 **Kabaghe AN**, ***et al*.**. Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural malawi. *PLoS ONE* 2017; **12**: e0172266.

34 **Chipeta MG**, ***et al*.**. Adaptive geostatistical design and analysis for prevalence surveys. *Spatial Statistics* 2016; **15**: 70–84.

35 **Thompson SK, Seber GAF.** *Adaptive sampling*. New York, NY: John Wiley and Sons, 1996.

36 **Bryce J, Steketee R.** Continuous surveys and quality management in low-income countries: a good idea. *The American Journal of Tropical Medicine and Hygiene* 2010; **82**: 360; author reply 361–362.

37 **Rowe AK.** 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* 2009; **80**: 971–979.

38 **Turner AG.** Sampling frames and master samples, 2003. (https://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/doc s/no_3.pdf).

39 **Bryce J**, ***et al*.**. 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* 2004; **94**: 406–415.

40 **Victora CG**, ***et al*.**. Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. *Lancet* 2011; **377**: 85–95.

41 **Ye Y**, ***et al*.**. Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa?. *BMC Public Health* 2012; **12**: 741.

42 **Byass P**, ***et al.***. Lessons from history for designing and validating epidemiological surveillance in uncounted populations. *PLoS ONE* 2011; **6**: e22897.

43 **Jha P.** Counting the dead is one of the world's best investments to reduce premature mortality. *Hypothesis* 2012; **10**: 1–6.

44 **Alkema L, Raftery A, Brown T.** Bayesian melding for estimating uncertainty in national hiv prevalence estimates. *Sexually Transmitted Infections* 2008; **84**(Suppl. 1): i11–i16.

45 **Alkema L, Raftery A, Clark S.** Probabilistic projections of hiv prevalence using Bayesian melding. *The Annals of Applied Statistics* 2007; **1**: 229–248.

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

47 **Kahn K**, ***et al.***. Profile: Agincourt health and socio-demographic surveillance system. *International Journal of Epidemiology* 2012; **41**: 988–1001.

48 **Kahn K**, ***et al.***. 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* 2007; **35**(69 Suppl.): 8–20.

49 **Clark SJ**, ***et al.***. Cardiometabolic disease risk and hiv status in rural South Africa: establishing a baseline. *BMC Public Health* 2015; **15**: 135.

50 **Clark SJ**, ***et al.***. Young children's probability of dying before and after their mother's death: a rural South African population-based surveillance study. *PLoS Medicine* 2013; **10**: e1001409.

51 **Gómez-Olivé FX**, ***et al.***. Prevalence of hiv among those 15 and older in rural South Africa. *AIDS Care* 2013; **25**: 1122–1128.

52 **Houle B**, ***et al.***. The unfolding counter-transition in rural South Africa: mortality and cause of death, 1994–2009. *PLoS ONE* 2014; **9**: e100420.

53 **Houle B**, ***et al.***. Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status. *International Journal of Epidemiology* 2013; **42**: 1444–1454.

54 **Kabudula CW**, ***et al.***. Assessing changes in household socioeconomic status in rural South Africa, 2001–2013: a distributional analysis using household asset indicators. *Social Indicators Research* 2017; **133**: 1047–1073.

55 **Diggle PJ, Tawn JA, Moyeed RA.** Model-based geostatistics (with discussion). *Applied Statistics* 1998; **47**: 299–350.

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

57 **Rue H, Martino S, Chopin N.** Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**: 319–392.

58 **Denison D, Holmes C.** Bayesian partitioning for estimating disease risk. *Biometrics* 2001; **57**: 143–149.

59 **Wakefield J, Simpson D, Godwin J** Comment: getting into space with a weight problem. discussion of, "model-based geostatistics for prevalence mapping in low-resource settings", by P. J. Diggle and E. Giorgi. *Journal of the American Statistical Association* 2016; **111**: 1111–1119.

60 **Diggle PJ, Menezes R, Su T-L.** Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2010; **59**: 191–232.

61 **Lohr SL.** *Sampling: design and analysis* (2nd edition). Boston, MA: Brooks/Cole, 2010.

https://doi.org/10.1017/gheg.2017.15 Published online by Cambridge University Press