Planning a Sample Registration System

Sampling Designs & Data Use

Agbessi Amouzou, PhD

Johns Hopkins University

September 8, 2025

Outline

- Why care about sampling design in an SRS?
- Elements of a sampling design
- Sample size
- Sampling procedures and weighting
- Other considerations
- Data use
- Example of Zambia – (Stephen)

SAMPLING DESIGNS

Why care about sampling design in an SRS?

1. Ensure representativeness of the SRS sample at national and subnational (provinces, districts) levels
    - Probability of everyone being in the sample is known
2. Estimate rigorously the precision or confidence intervals around indicators derived from the SRS data
3. Ensure indicators derived are credible, reliable, and can be trusted
4. Ensure transparency in the selection of the population included
5. Promote replicability and reproducibility
6. Compare estimates from an SRS to other data sets

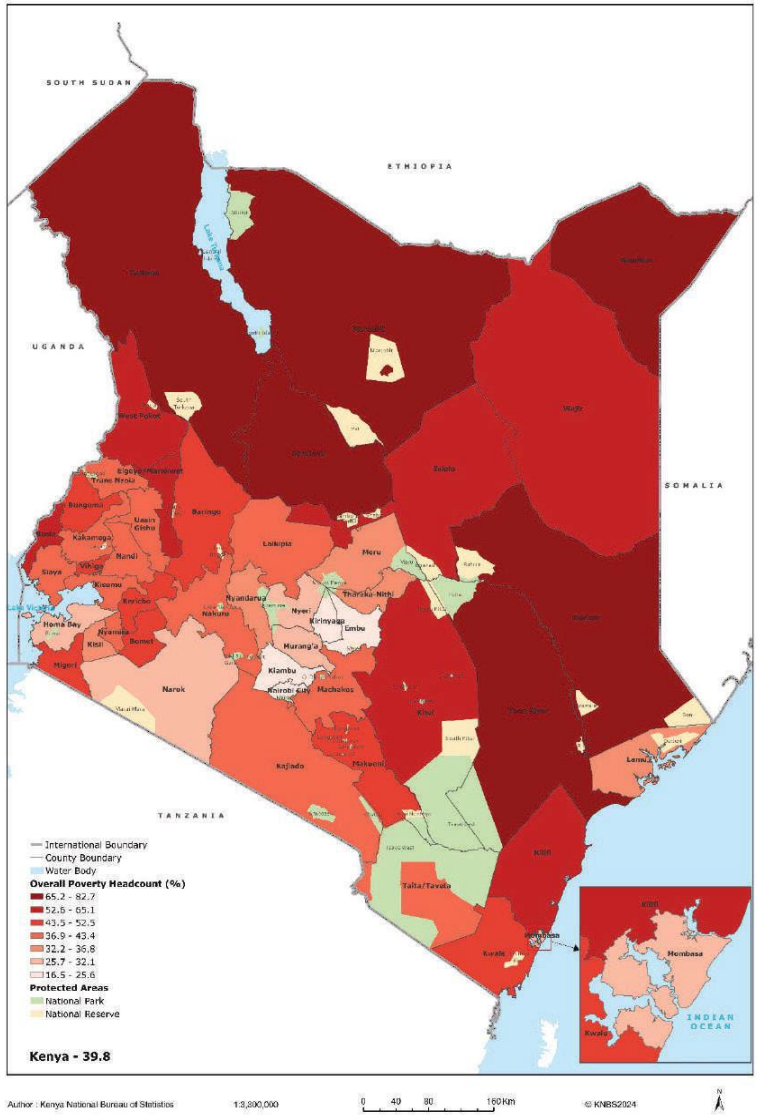
Elements of a Sampling Design

1. Understanding of the government health priorities and potential users of the SRS data
2. Level of representativeness of the sample (statistical domains)
   - National, regions/provinces, districts, urban/rural?
3. Sample strata, i.e., groups within which an independent sample is drawn
   - e.g., regions by urban/rural
4. Smallest geographic sampling units (or primary sampling units or clusters)
   - E.g., Census EAs, villages, districts, subdistricts
5. Complete sampling frame of clusters, organized by statistical domains (often available at the National Statistics Office)
   - Census frame recommended as it is often complete and is often updated during a population census
6. Sample size
7. Sample selection procedures and sample weights

1. Government health priorities and potential users of the SRS data

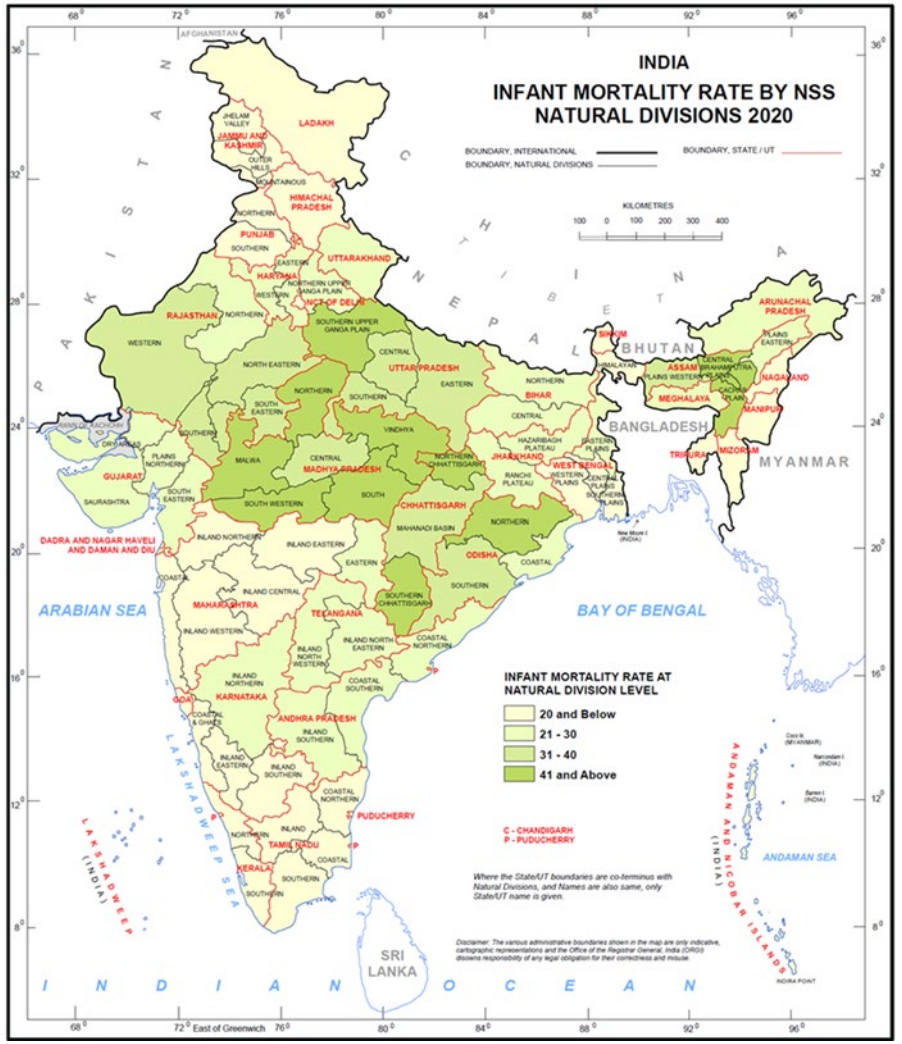
- Understand government priority subnational areas that require frequent and more precise data for targeted health interventions or performance assessment
  - E.g., From national health strategic plans
  - Involve MoH and other sectors
- Who are the other direct data users and what level data granularity they want like to see
  - Partners implementing large multi-year health programs
- Potential linkage with other existing data

Overall poverty level at county level, Kenya 2022



2. Level of representativeness of the sample (statistical domains)

- Statistical domains are the lowest geographic areas at which precise indicators will be generated
  - E.g., India uses Natural divisions
- It is determined through consultation with the government and stakeholders
- Typically uses subnational administrative areas
  - E.g. regions/provinces, urban/rural, districts
- Some subnational areas may be oversampled
  - E.g. oversample a poor region to produce district level estimates
- More statistical domains you create, the higher your sample size and your cost



3. Sample strata

- Groups of sampling units within which an independent sample is drawn
  - E.g. rural/urban areas area in each region, special populations/settlements
- Stratification can be used to ensure representation of some groups in your sample, especially minority groups that may not be picked in a non-stratified sample
- Statistical domains may be used as strata
- Stratification creates unequal probability samples requiring use of sampling weights (see next slides)

4. Smallest geographic sampling units (clusters)

- Smallest sampling units within which a complete surveillance is established
- Must be decided based on the availability of the complete roster
- Population census enumeration areas (EAs) offer more complete and rigorous sampling units
  - Come from the recent population census
  - Often 100-150 households
  - Digital and sketch maps are often available at the National Statistical Office
  - Covers the entire country and is stable
  - However, it can cut across villages or communities
- Subnational areas, such as villages/blocks, districts/communes, subdistricts, may be used but can be unstable over time
- Must have clearly identifiable boundaries

5. Complete sampling frame of clusters

- A complete roster of clusters (e.g. EAs) organized by administrative units, statistical domains, strata, and population
- Available with the National Statistical Office, with maps and, in some cases, geocoordinates of boundaries and key landmarks
    - Will need to work with NSO as the roster may not be accessible to everyone
- Does not need to be entirely up to date on population size but must cover entire sampling area
- Avoid sampling frame that changes too quickly (e.g. list of villages, districts)

Region	District	Village/commune	Id Enumeration area	# Households	Populatino
1	11	111	1111111	150	750
1	11	112	1111112	125	700
1	11	113	1111131	90	500
1	11	113	1111132	100	500
1	11	114	1111141	120	600
…	…	…	…	…	…

6. Sample size

Drivers of sample size:

1. Indicators to measure
2. Statistical precision and domains
3. Period of measurement for mortality indicators
4. Government health priorities
5. Budget

6. Sample size: choice of indicators

- Mortality indicators
  - ✓**Crude death rate** = annual deaths / mid-year population
  - ✓**Infant mortality rate** = infant deaths (T) / live births (T)
  - ✓**Under-five mortality rate** = under-five deaths (T) / live births (T)
  - ✓**Adult mortality rate** = annual deaths (15-64) / mid-year adult population
  - ✓**Maternal mortality ratio** = maternal deaths (T) / live births (T)
  - ✓ **Cause-specific mortality rate** = deaths by cause (T) / target population (T)
  - ✓Others…
- Choose the indicator likely to generate sample sizes that will be enough to measure most other indicators, but not too unrealistically large
  - E.g., IMR
- Once the sample size is computed, calculate the precision of other indicators
- Sample size must be computed for each statistical domains

6. Sample size calculation

- m = mortality rate (in proportion)
- d = absolute margin of error, ie. Confidence interval is m ± d
- R = household response rate
- Deff = Design effect (borrow from existing surveys, e.g. DHS)
- N is the target population corresponding to the indicator denominator (e.g., N is #births if using IMR)
- N must be converted into the corresponding total population and number of households

$$N = Z_{\alpha/2}^2 * \frac{m * (1-m)}{d^2 * R} * Deff \quad \Longrightarrow \quad d = SQRT(Z_{\alpha/2}^2 * \frac{m * (1-m)}{N * R} * Deff)$$

6. Sample Size: Example

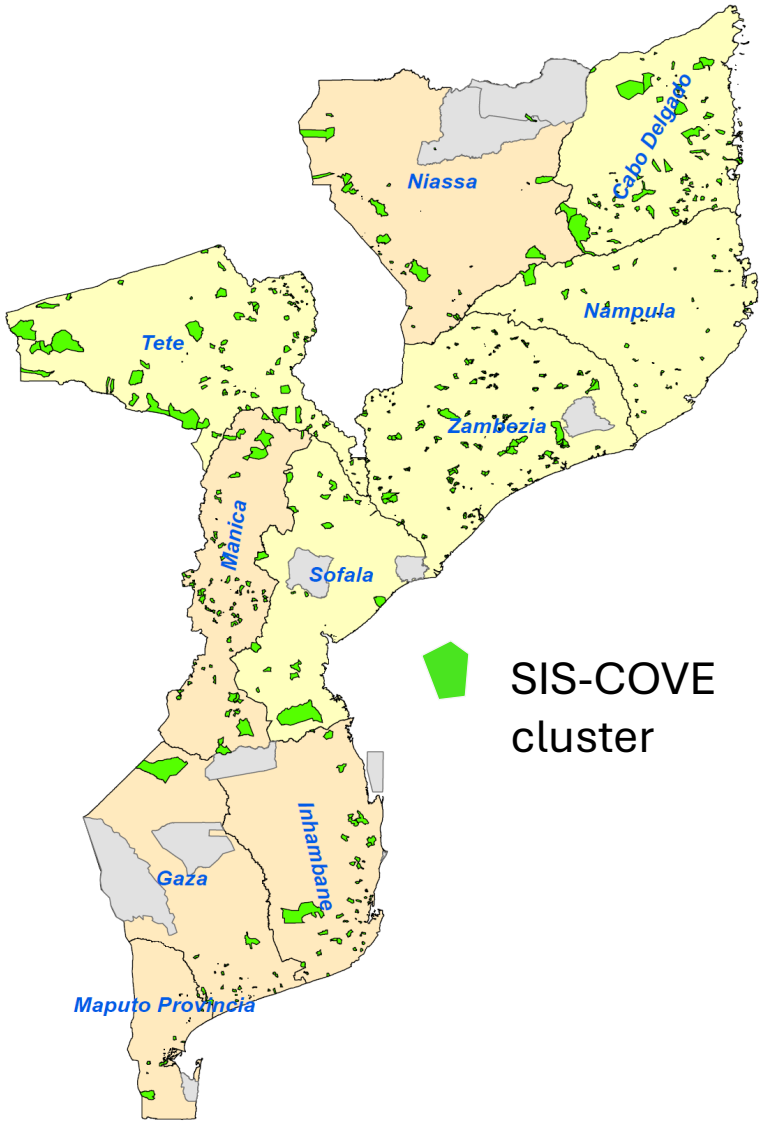
Domains	IMR (per 1000 LB)	Relative margin of error (2SE/M)	Absolute margin of error	Estimated # Births (Annual)	Crude birth rate (per 1000)	Average household size	Corresponding Number of households	Corresponding population	Number sampling units (assuming 125 HH per unit)	Estimated number infant deaths
Domain 1	100	**12%**	12.0	**3,961**	40	5	19,805	99,023	**158**	396
Domain 2	90	**12%**	10.8	**4,450**	35	5	25,428	127,141	**203**	400
Domain 3	50	**20%**	10.0	**3,010**	30	5	20,069	100,344	**161**	151
Domain 4	80	**12%**	9.6	**5,061**	40	5	25,306	126,530	**202**	405
Domain 5	150	**10%**	15.0	**3,591**	45	5	15,961	79,806	**128**	539
TOTAL				**20,074**			**106,569**	**532,844**	**853**	**1,891**

Assumes Design effect = 1.5; Non-response rate =10%

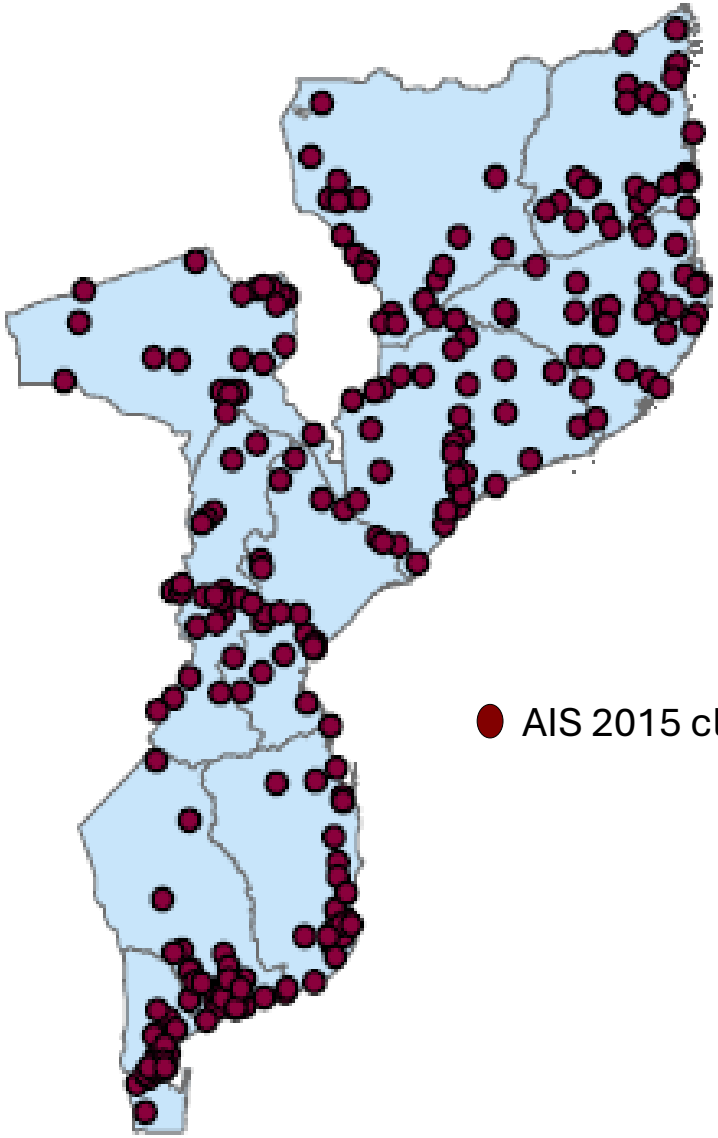
7. Sample selection procedures and design weights

- Select the sample within each stratum identified
- Organize sampling frame by administrative units (implicit stratification)
- Select primary sampling units (clusters) randomly with probability proportionate to size
  - Systematic random sampling
- Plot the sampled clusters on a map and verify that the sampling is acceptable
- Compare with other existing data as needed

Mozambique SIS-COVE Sample compared to National AIS 2015 Survey



SIS-COVE cluster



AIS 2015 cluster

7. Design weights: Inverse of probability of selection within sample stratum

- When all sample units don't have the same probability of selection, a design weight is needed to adjust the sample during analysis to obtain representative estimates

	Strata			National
	1	2	3	
Total Population	P1	P2	P3	P = (P1+P2+P3)
Sample Pop	S1	S2	S3	S=(3*S)
Probability of selection	S1/P1	S2/P2	S2/P3	

S is not always representative of P

Sample weights	P1/S1	P2/S2	P3/S3	

7. Design weights

- Sampling procedure: single stage stratified cluster sampling with probability proportionate to population size
  - Cs = total number of clusters in Stratum S
  - Ns = number of clusters selected in stratum S
  - Clusters numbered from i to Cs in stratum S
  - Cis = population size of cluster I stratum S
  - Ms = total population in Stratum S

$$Probability\ of\ selection\ of\ cluster\ i\ in\ stratum\ S = Pis = Ns * \frac{Cis}{Ms}$$

$$Design\ Weight\ for\ cluster\ i\ in\ stratum\ S = Wis = \frac{1}{Pis}$$

- Design weights can be adjusted for non-response at the cluster and household levels
- It is critical to document the sampling procedures to allow computation of design weights

Who to involve in the sample design?

- Sampling experts from the National Statistical Office
  - Compute the sample size
  - Draw the sample
  - Calculate sample weights
- Ministry of Health and other stakeholders
  - For input on the distribution of the sample
- Demographer, statistician, or Epidemiologist

Resources on Viva

- Visit viva.jhuhost.org to access the repository of tools that help with sampling design

Homework for countries planning an SRS

Complete this table to summarize your sampling design

Activities		Describe	Who has been involved
1	Government health priorities and potential users of the SRS data		
2	Level of representativeness of the sample (statistical domains)		
3	Sample strata		
4	Smallest geographic sampling units (clusters)		
5	Availability of complete sampling frame of clusters		
6	Sample size		
7	Sample selection procedures and sample weights		

Data Use

Utility of SRS Data

1. Understanding and monitoring health trends and disparities
   - ✓ All-cause mortality trends
   - ✓ Leading cause of death
   - ✓ Disease burden
   - ✓ Improved country-specific health research
   - ✓ Disaggregation by age, sex, geography, and SES
2. Generating evidence to inform priority interventions
   1. Population at highest risk (e.g., newborn, maternal, senior)
   2. Disparities
   3. Resource allocation
3. Assessing the effectiveness of policies and programs
   1. Strong evaluation designs (pre/post with comparison)
   2. Impact evaluations
4. Serving as an early warning for crises
   1. Pandemics
   2. Emergency response during disasters
   3. Unusual patterns of mortality
5. Linking and assessing other data systems
   1. CRVS, HMIS, Others

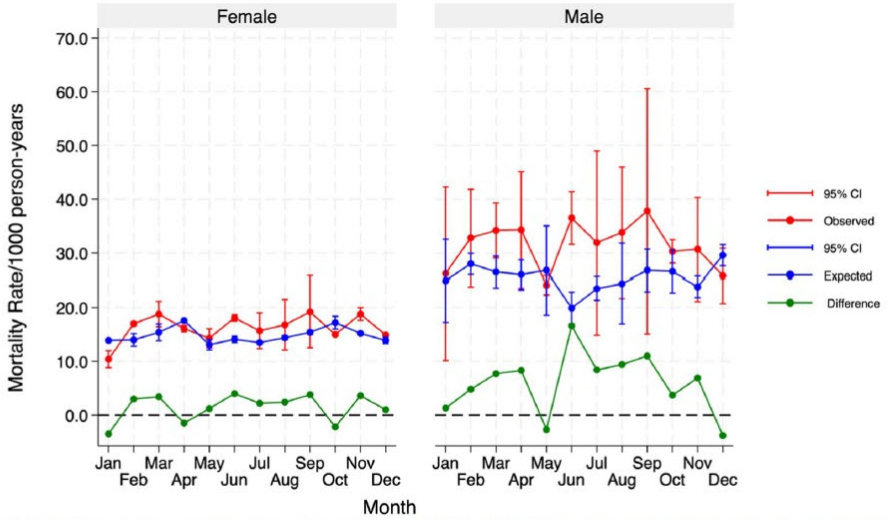
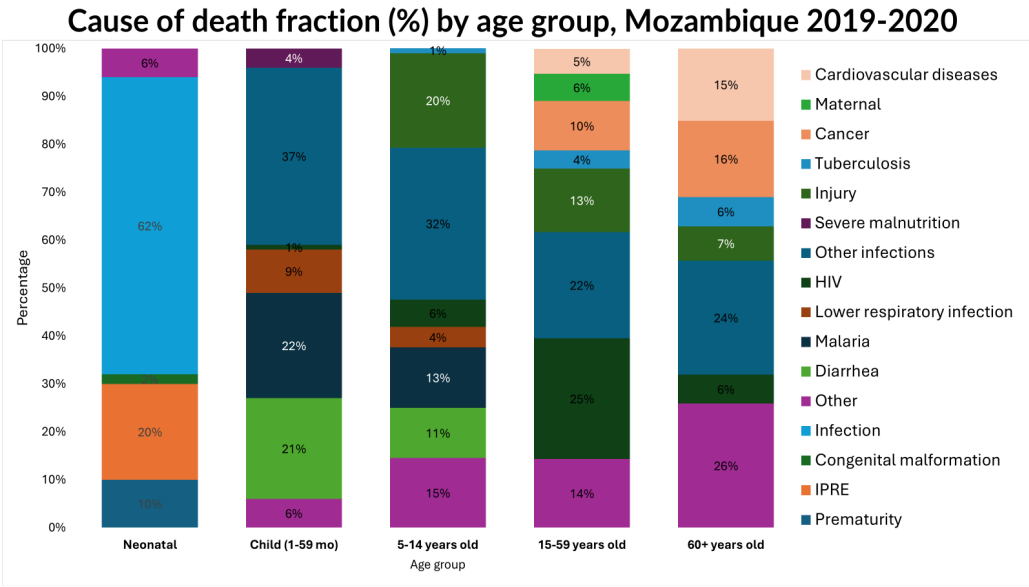


Fig. 4 Monthly trends in excess mortality by gender during the COVID-19 pandemic (2020–2021) (using outputs from the predictions of the Poisson generalised additive model) in the Navrongo HDSS study area

Azongo *et al. Population Health Metrics (2025) 23:31*
https://doi.org/10.1186/s12963-025-00389-7



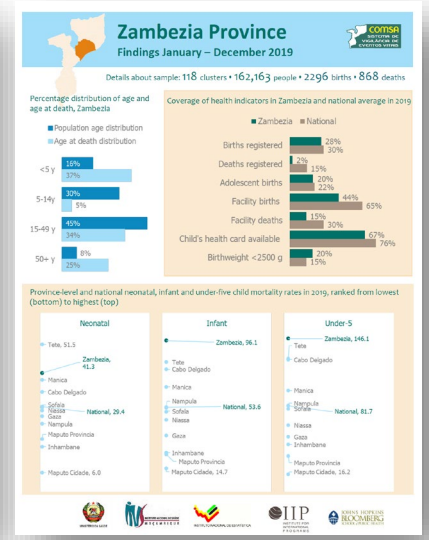
Macicame et al. 2023

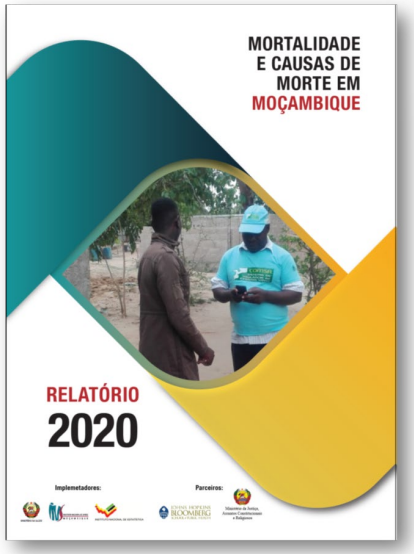
Who are the data users?

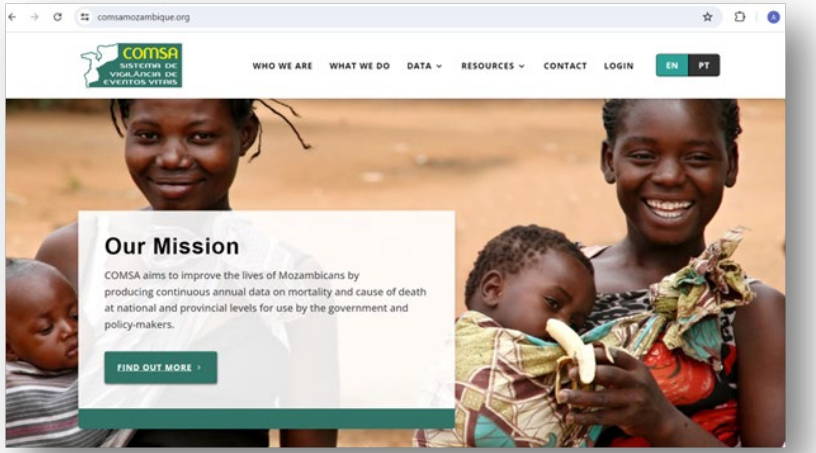
Utility	Users
1. Understanding and monitoring health trends and disparities	Ministry of Health Departments, Other sectors (e.g. nutrition, education, gender and social protection); Subnational health officials; partners; Public health research and statistics institutions; Health Professional Associations; global health community
2. Generating evidence to inform priority interventions	Ministry of Health decision-makers; subnational health officials; funders;
3. Assessing the effectiveness of policies and programs	Health program managers; Funders; Technical partners; subnational health officials
4. Serving as an early warning for crises	Public Health Institutes; Department of Public Health; Humanitarian health office
5. Linking and assessing other data systems	CRVS, HMIS,

What platforms for data dissemination and sharing?

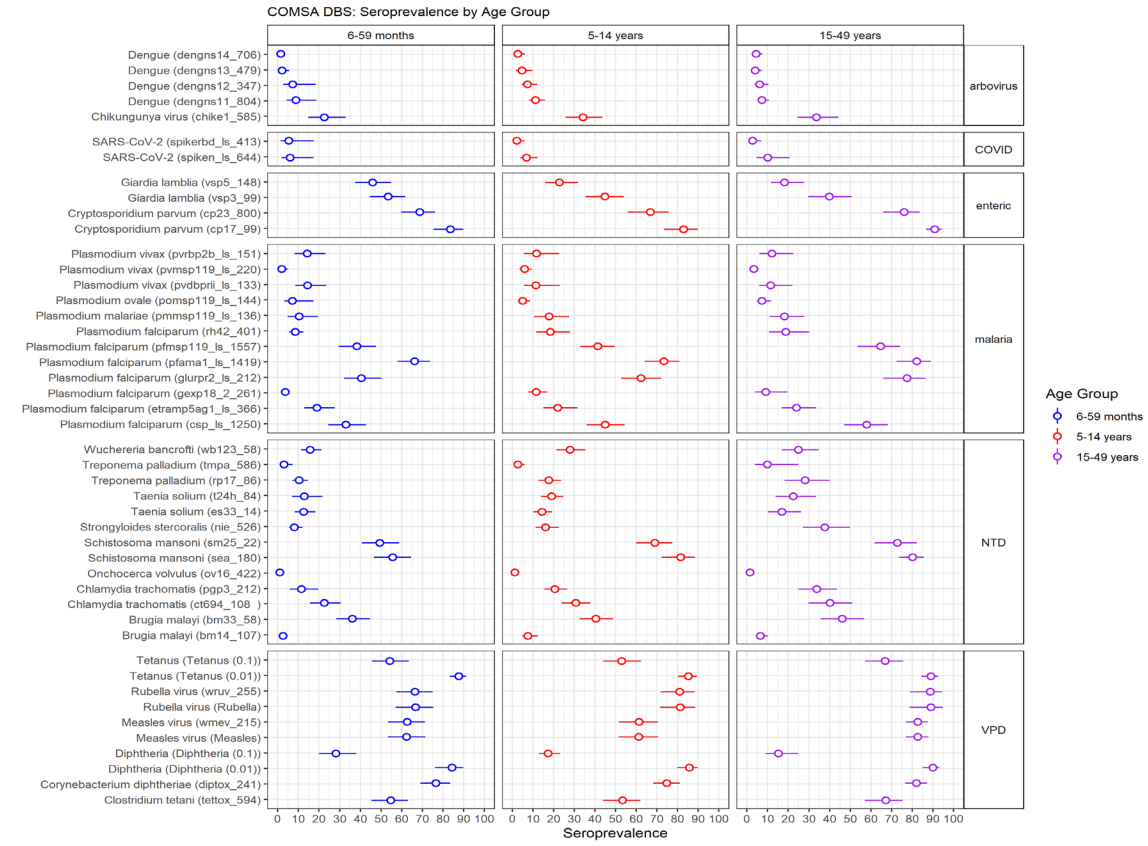
1. Annual data reports sharing detailed results
2. Synthesis report and facts sheets at national and subnational levels
3. Specific national groups (e.g. National Advisory Group; MoH Leadership Groups)
4. Web platforms with public data access modalities (website with updated and interactive data visualization)
5. Linkage with other systems' platform (e.g. DHIS-2)
6. Peer-reviewed publications

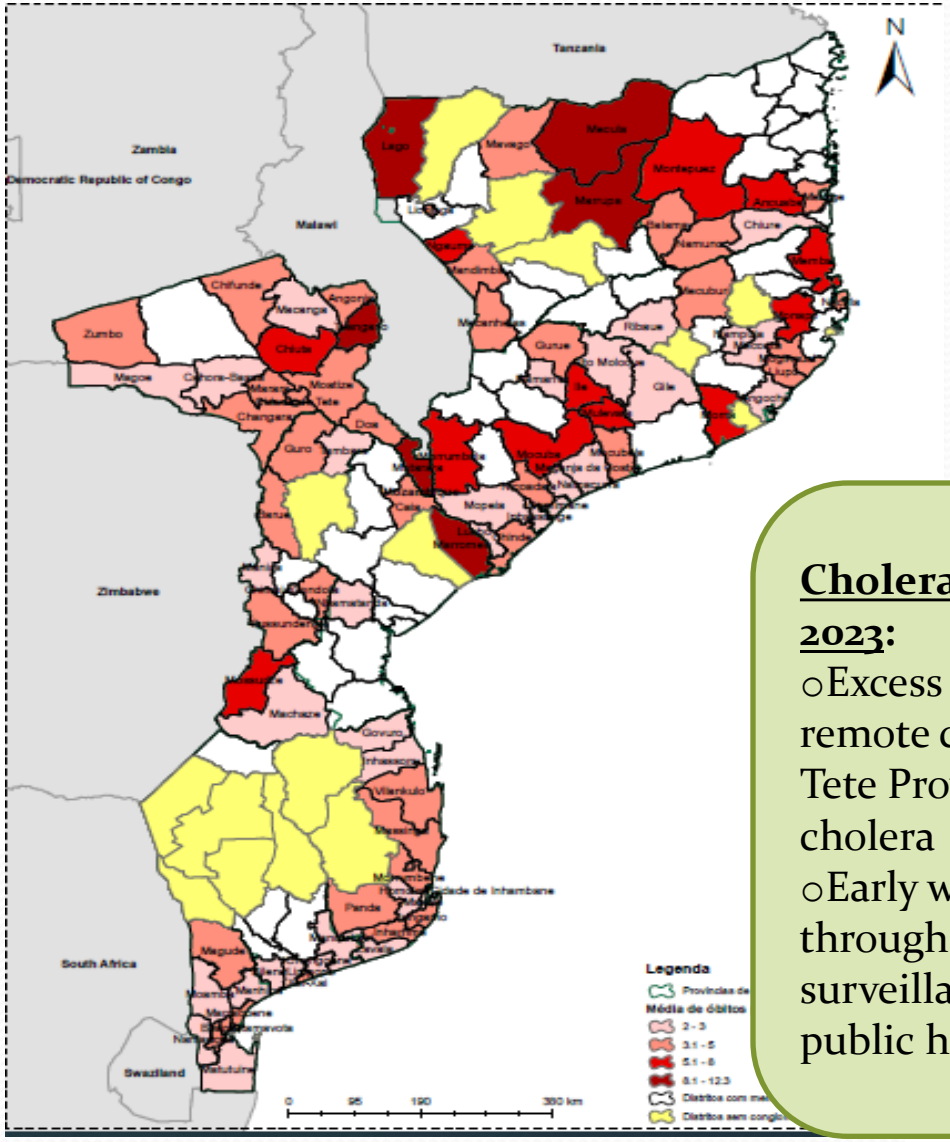






Leveraging SIS-COVE to improve community-based disease surveillance and outbreak investigation





<u>VPD Seroprevalence in Zambezia</u>:

- Diphtheria: 87.2% (95%CI 84.3-89.6)
- Tetanus: 86.9 (95%CI 83.6-89.6)
- Rubella: 83.0% (95%CI 78.3-86.8)
- Measles: 69.1% (95%CI 63.3-74.4)

<u>Cholera outbreak in 2023</u>:

- Excess mortality in a remote community in Tete Province due to cholera
- Early warning system through mortality surveillance to detect public health threats

Mozambique: SIS-COVE Mortality data being used to increase general literacy and for policy decision making in different sectors

SIS-COVE and DHS as the main sources of data to inform the development of the 2025-2029 Health Sector Strategic Plan

REPÚBLICA DE MOÇAMBIQUE

MINISTÉRIO DA SAÚDE

Direcção de Planificação e Cooperação

Plano Estratégico do Sector da Saúde

PESS 2014-2019

(Extensão 2020-2024)

«O NOSSO MAIOR VALOR É A VIDA»

Maputo, Agosto 2022

Meetings

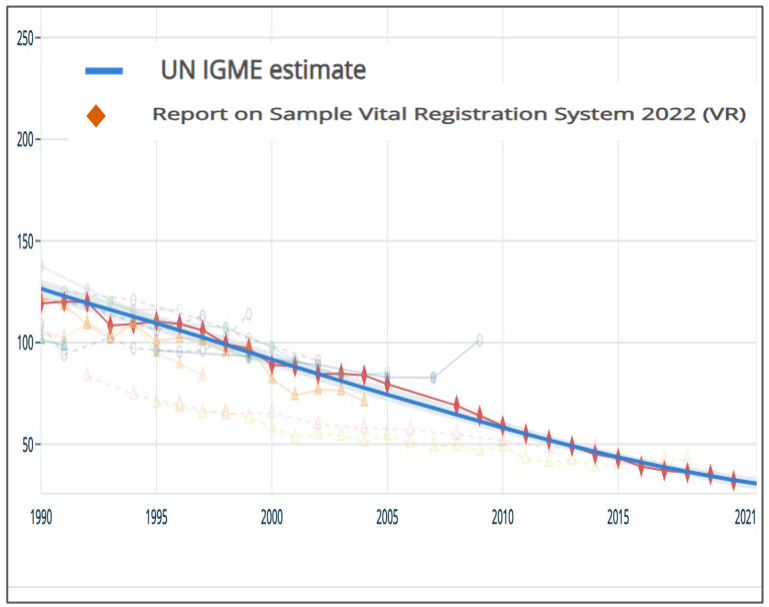
Radio

Newspapers

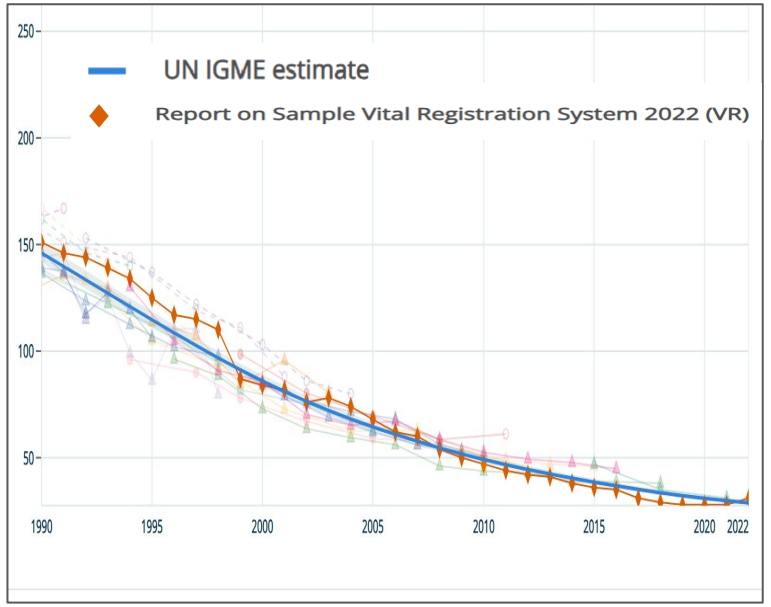
Live interviews

SRS Data Drive Child Mortality Estimates by UN-IGME

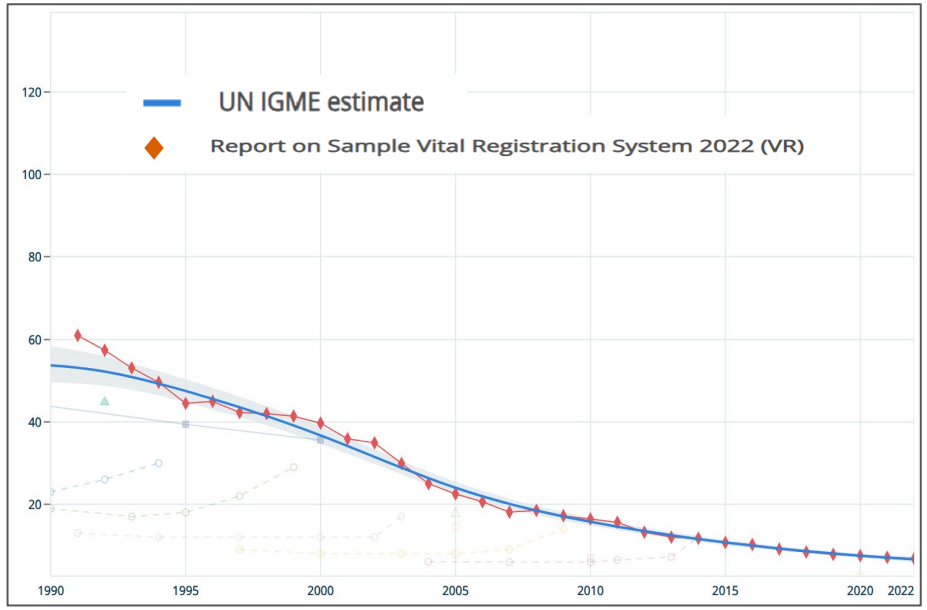
India



Bangladesh



China



- When available and well functioning, SRS can be one of the best sources of mortality data

Be aware of barriers to data use

1. Untimeliness of the results
2. Data quality and lack of transparency on how data and results are obtained
3. Limited linkage or interoperability with other systems
4. Low capacity for data analysis and interpretation
5. Ethical and legal concerns about data privacy

Thank You!

Minimum Data in a SRS

Numerators:

- Number of deaths and its distribution by age, sex, geography, and cause of death

Denominators:

- Births (stillbirths, perinatal, childhood mortality rate)
- Population by age and sex (person-years lives or population at risk)

-> An SRS must at minimum collect data on population by age and sex, pregnancy outcomes, deaths, and causes of death

Infant mortality rate

$$IMR = \frac{Infant\ deaths\ (T)}{Births\ (T)}$$

Age specific mortality rate

$$_nM_x = \frac{_nD_x(T)}{_nPYL_x(T)}$$

PYL: Person-year lived

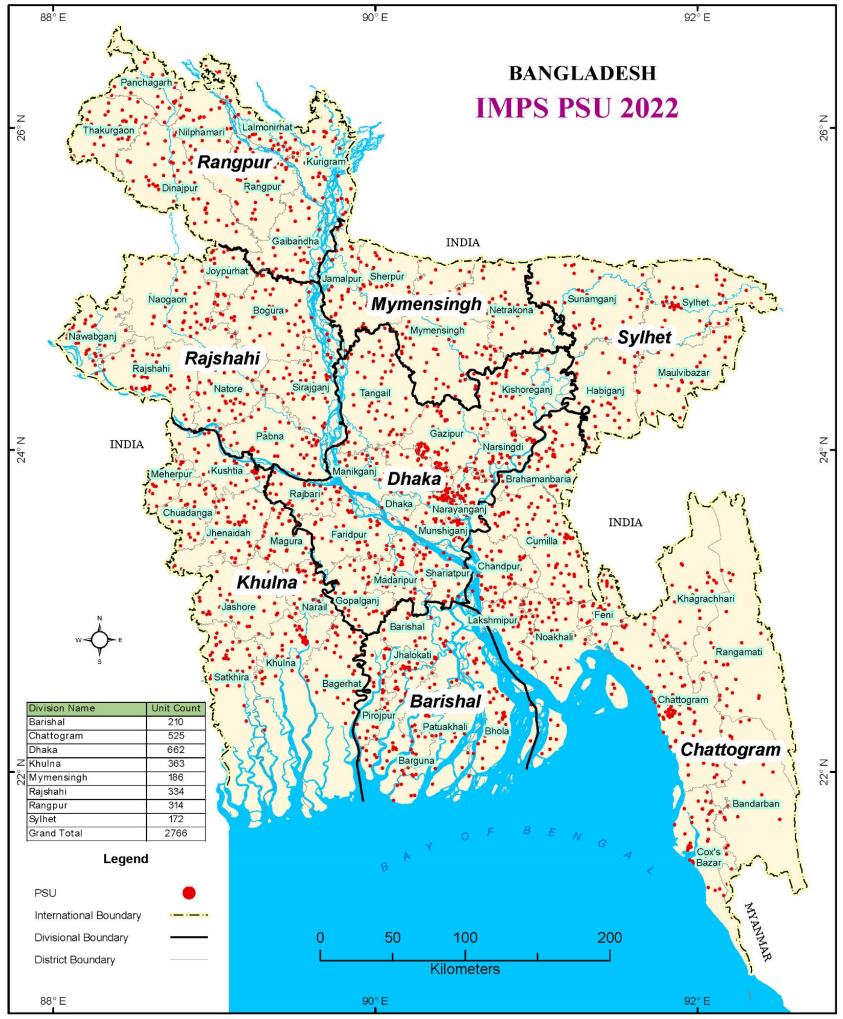
$_nM_x$ = mortality rate between ages x and x+n

Age specific probability of death

$$_nq_x = \frac{n * _nM_x}{[1+(n-s) * _nM_x]}$$

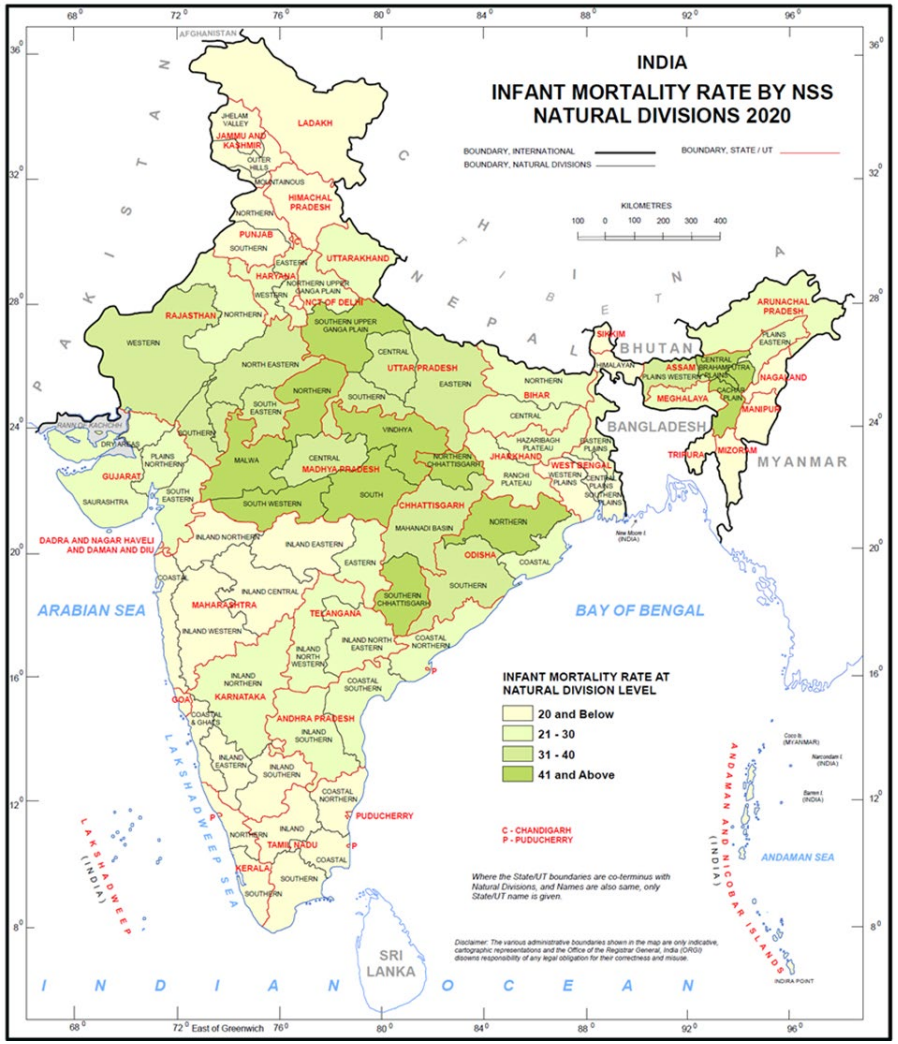
Current sampling experiences

Bangladesh



- Started in 1980
- Sampling frame: 2022 Census EAs
- Sample size in 2024:
  - 2,766 EAs
  - 313,140 households
  - >1.3 million population
- Sampling method
  - Statistical domains and strata: 7 divisions X urban/rural/city corporation
  - Single stage stratified random sampling of EAs with PPS
- Sample represents 0.8% of the pop

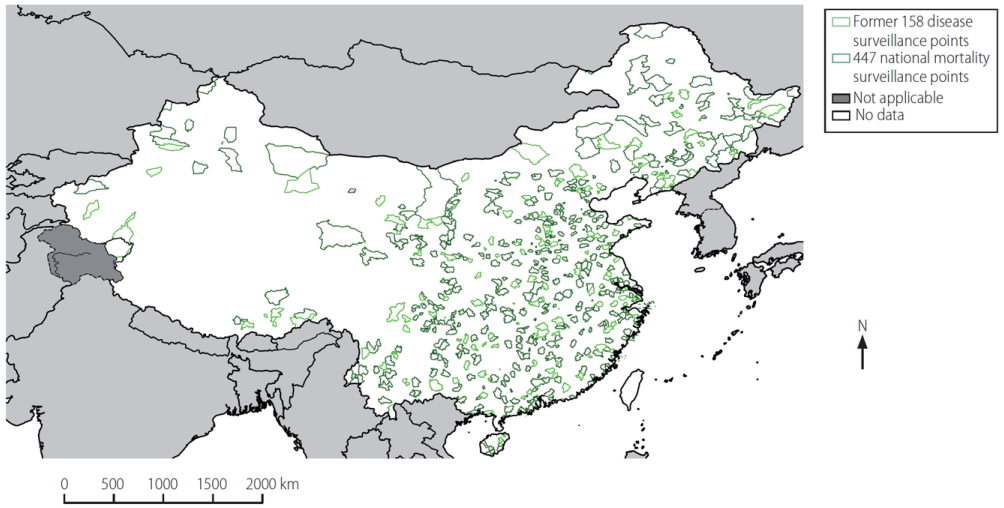
India



- Started: 1968-70
- Sampling frame: 2011 census EAs
- Sample size based on IMR at natural division level (10% relative margin for large states and 15% for smaller states)
- Sample: 8853 EAs (2014)
- Sampling domains: Natural divisions or small states
- Sample selection: Mix of single and two-stage stratified random sampling (depending on size of strata)
- Sample represents 0.7% of total population

China

Fig. 2. **Surveillance points, national mortality surveillance system, China, 2013**



Note: Dark green lines represent 447 counties with national mortality surveillance points; light green lines represent 158 counties which also have disease surveillance points. The map was produced using QGIS version 2.8.3 (QGIS Development Team).
Source: Map data for China from the Natural Resources and Geospatial Base Information Database (National Development and Reform Commission, China). Map data for neighbouring countries from WHO.

- Integration of VS and Disease surveillance points in 2013
- Sampling frame: 2010 pop census
- Cluster: counties/districts
- Sample: 605 clusters determined based on expert consultation
- Sampling procedure: Iterative selection controlling for selected demographic and socio-economic indicators to ensure representativeness
- Sample represents 24% of total population

Utility of an SRS for other Surveillance Systems

Existing Systems	Roles of SRS
CRVS	▪ Evaluate completeness ▪ Promote registration of vital events ▪ Facilitate targeted implementation research studies to support rapid scale-up
HMIS	▪ Complement with community data ▪ Understand the cause of death at the community level ▪ Understand the profile of non-facility users ▪ Serve as a platform for evaluating and testing community-based approaches
Other special surveillance systems (e.g. MPDSR)	▪ Complement with community data ▪ Evaluate completeness and data quality