NTNU

Norwegian University of Science and Technology

Prevalence Mapping in Low- and Middle-Income Countries

Geir-Arne Fuglstad
Department of Mathematical Sciences, NTNU

Many people involved

I will touch on work that is a collaborative effort of many people:

- Jon Wakefield and his group (and alumni) at University of Washington
- Andrea Riebler, and former students/post.docs. at NTNU
- Peter Gao at San José State University
- Zehang Richard Li at University of California Santa Cruz
- UN Inter-agency Group for Child Mortality Estimation (UN IGME)

Parts are based on the forthcoming discussion paper:

- Jon Wakefield, Peter Gao, Geir-Arne Fuglstad, and Zehang Richard Li. (2025). The Two Cultures for Prevalence Mapping: Small Area Estimation and Model-Based Geostatistics. *Statistical Science*. In press.

Background

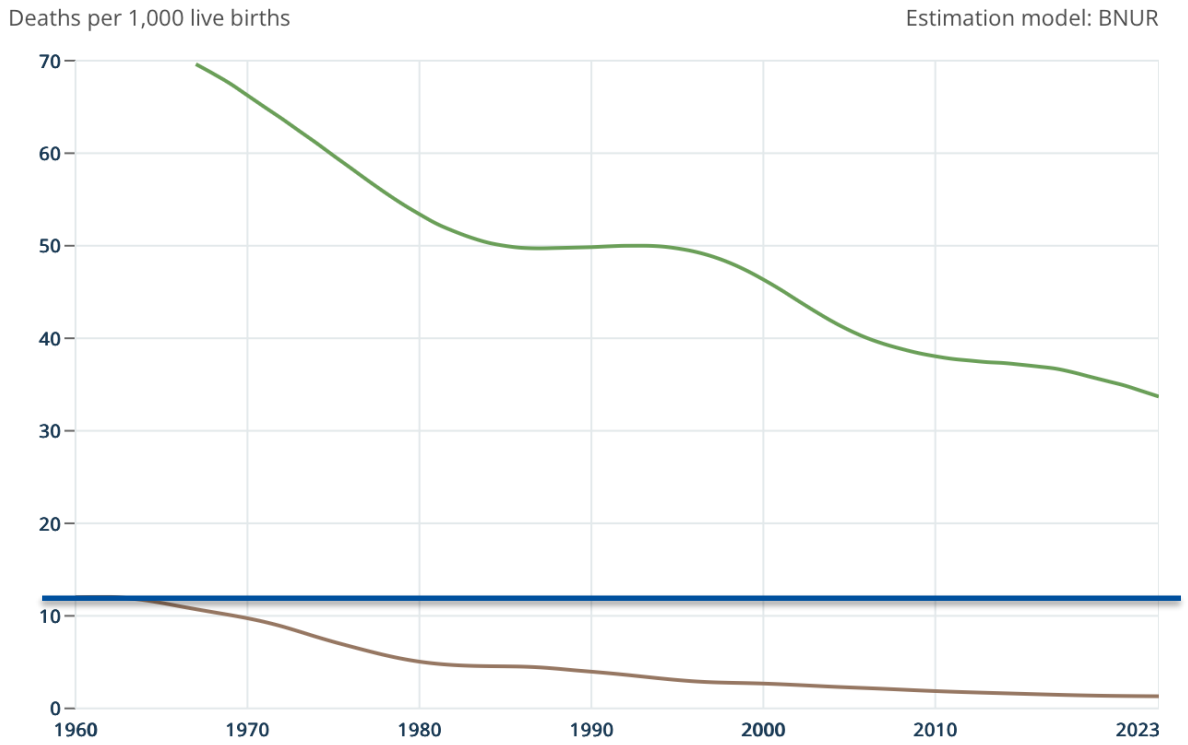
The United Nations (UN) has committed to work towards the Sustainable Development Goals (SDGs) with ambitious goals for 2030.

Global Health

Need to monitor progress globally.

Example:

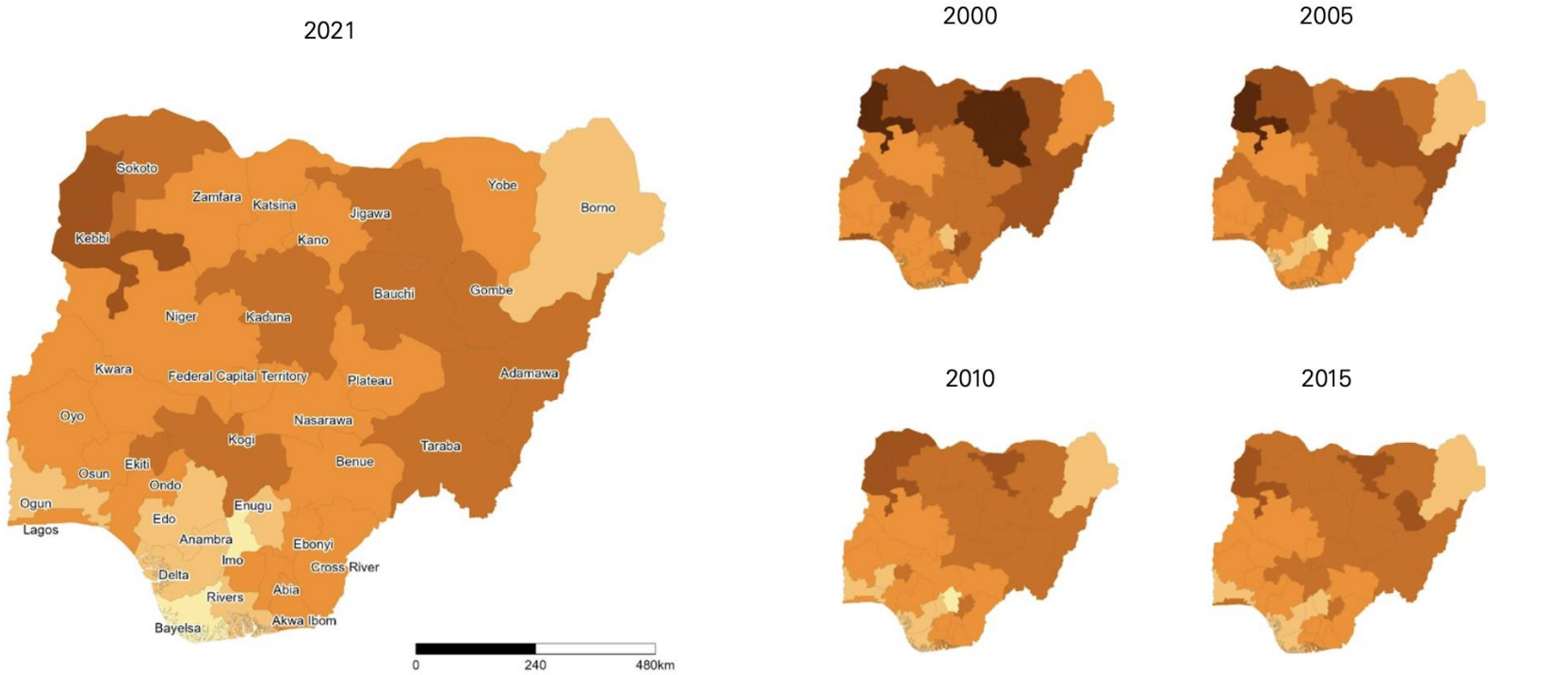
- SDG 3.2: Reduce NMR to 1.2%.
- Huge global differences!
  - Norway (brown): 0.1%
  - Nigeria (green): 3.4%
- Norway reached SDG 3.2 around 1960



Neonatal mortality rate (NMR):
Number of deaths within 28 days per live birth

Subnational hetrogeneity

«Leave No One Behind»



In 2000, NMR ranged from 26 to 65 deaths per 1,000 live births. By 2021, NMR ranged from 24 to 52 deaths per 1,000 live births.

The challenge is data

- Low- and middle-income countries (LMICs) often have deficient vital registration systems
- Low availability of auxilliary data
- Huge effort in conducting household surveys approximately every 5th year
  - More than 470 Demographic and Health Surveys (DHS) conducted across in over 90 LMICs
  - More than 350 Multiple Indicator Surveys (MICS) conducted across 120 countries

Example: Zambia 2018 DHS

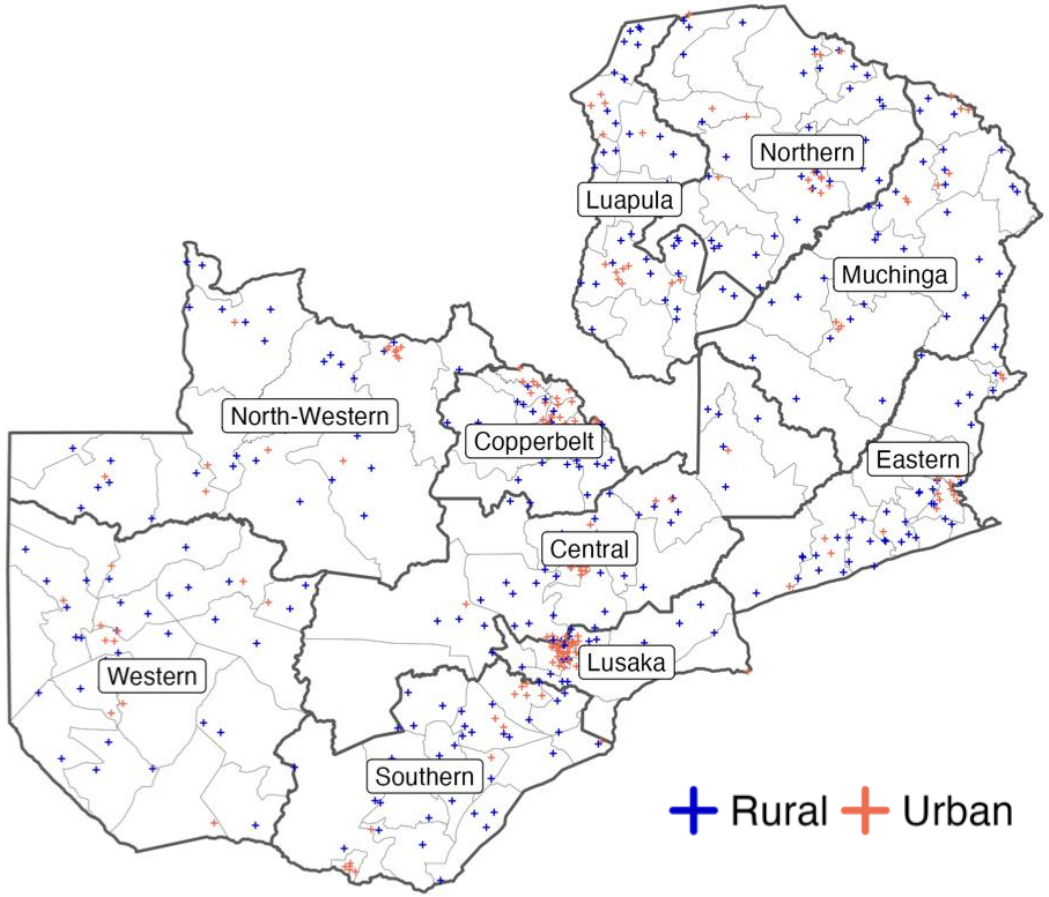
Consider prevalence of HIV among women aged 15—49 in Zambia in 2018.

Sampled under stratified two-stage clustering:

- 545 clusters (with GPS locations)
- 12893 women

Map shows

- 10 provinces (admin1 areas)
- 116 districts (admin2 areas).



National prevalence

- The population $U = \{1,2, \ldots, N\}$ are all N women aged 15—49 in Zambia
- Their HIV status is $y_i \in \{0,1\}$ for $i \in U$
- The target quantity, national prevalence, is then

$$p = \frac{\text{Number HIV positive women}}{\text{Number of women}} = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} 1} = \frac{\sum_{i \in U} y_i}{N}$$

Problem: Both numerator and denominator are unknown!

Direct estimates

Each woman has a probability π_i of being sampled.

When we sample $n < N$ women, $S \subset U$, the Hàjek estimator of prevalence is

$$\hat{P} = \frac{\sum_{i \in S} y_i/\pi_i}{\sum_{i \in S} 1/\pi_i},$$

where the scaling by $1/\pi_i$ accounts for unequal sampling probability.

We can also estimate the variance $\mathrm{V} = \mathrm{Var}[\hat{P}]$, but need to account for design.

This is called a **direct estimator**.

Subnational direct estimates

At subnational level (admin1 or admin2) we have areas $a = 1,2,\dots,A$ with:

- N_a: number of women
- $U_a \subset U$: subpopulation of N_a women
- $S_a \subset U_a$: subsample of $n_a < N_a$ women

This gives rise to subnational targets and direct estimators:

$$p_a = \frac{\sum_{i \in U_a} y_i}{\sum_{i \in U_a} 1} = \frac{\sum_{i \in U_a} y_i}{N_a} \quad \text{and} \quad \hat{P}_a = \frac{\sum_{i \in S_a} y_i/\pi_i}{\sum_{i \in S_a} 1/\pi_i}, \qquad \text{a} = 1,2,\dots,A.$$

Key point: Direct estimates can only use women sampled in the given area!

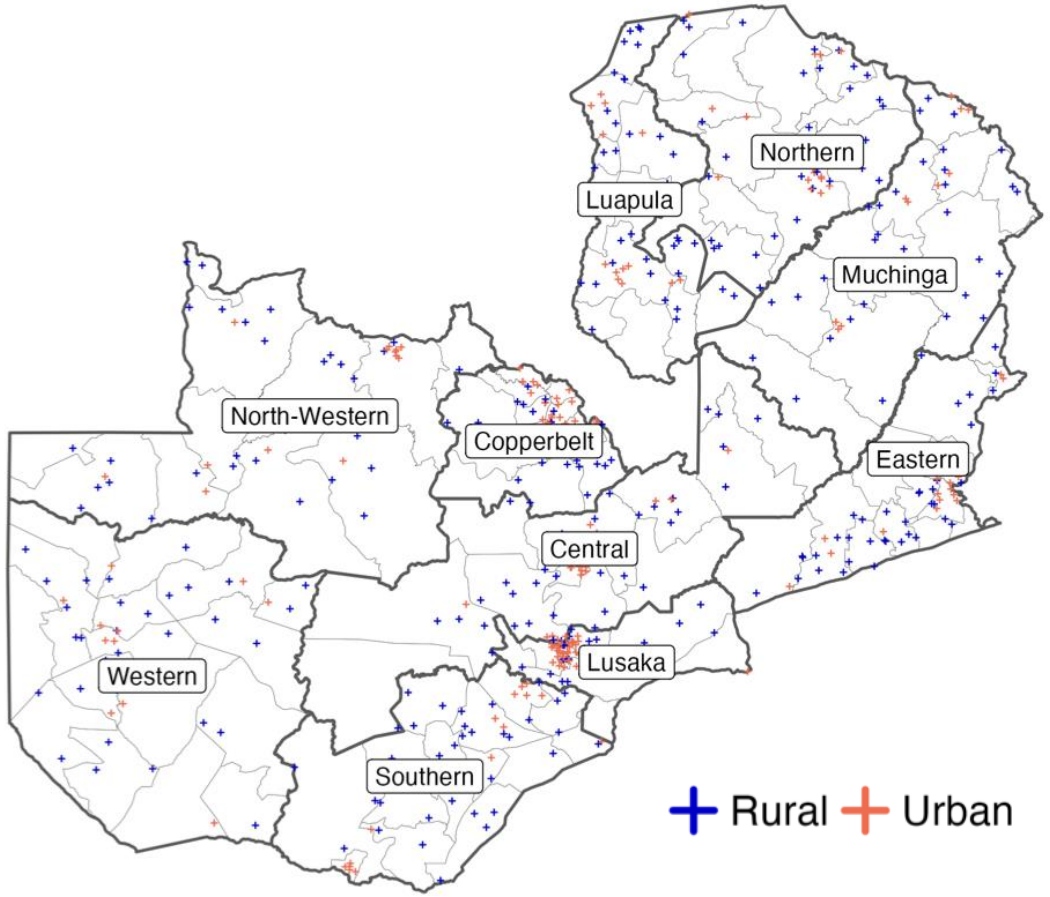
Data coverage

Admin1 areas were planned and have sufficient data.

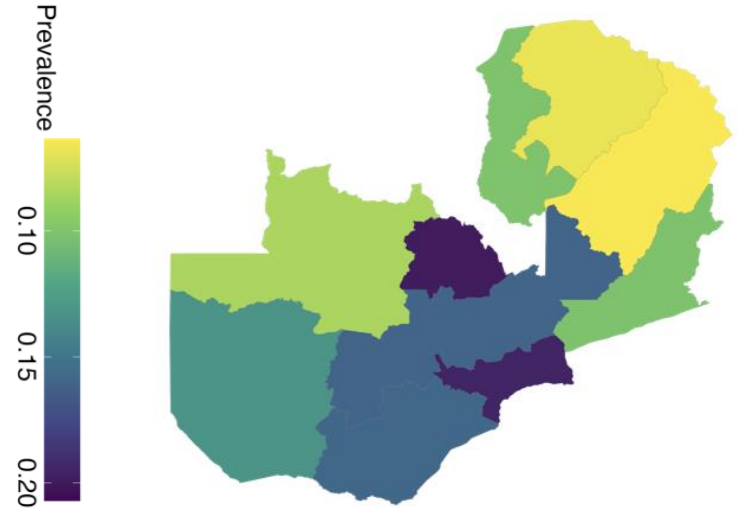
Admin2 areas were unplanned and data is very sparse.

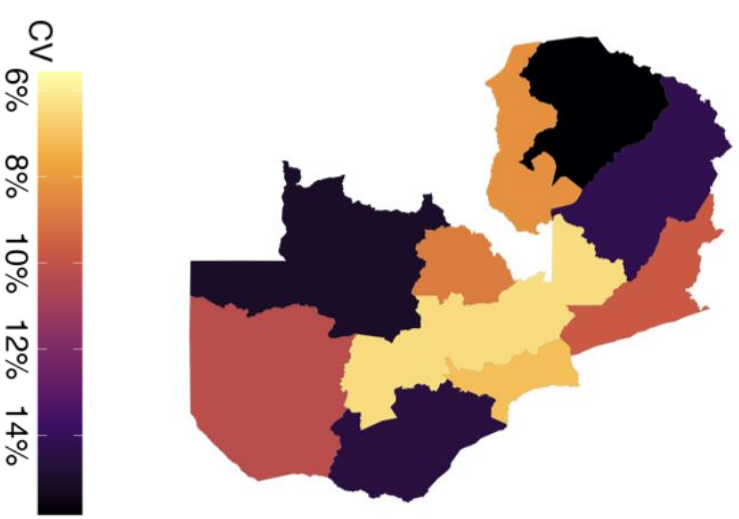
Often measure uncertainty through a coefficient of variation (CV)

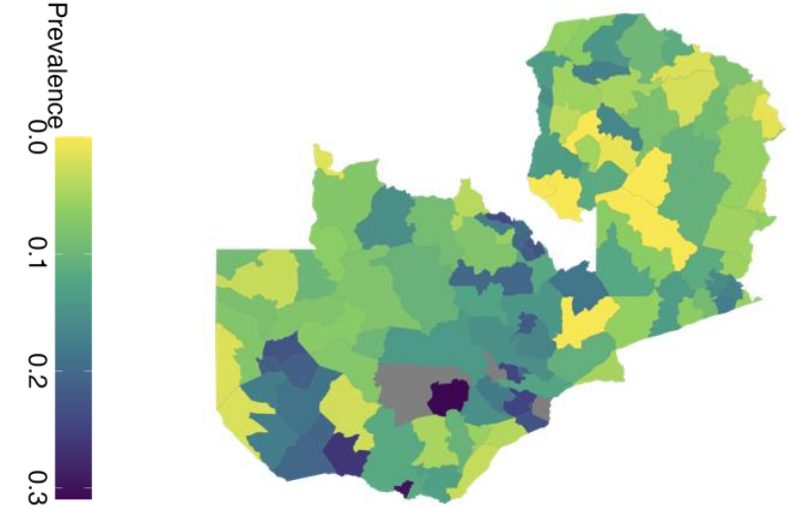
$$\mathrm{CV}_a = \frac{\mathrm{SD}[\hat{P}_a]}{\hat{p}_a}$$

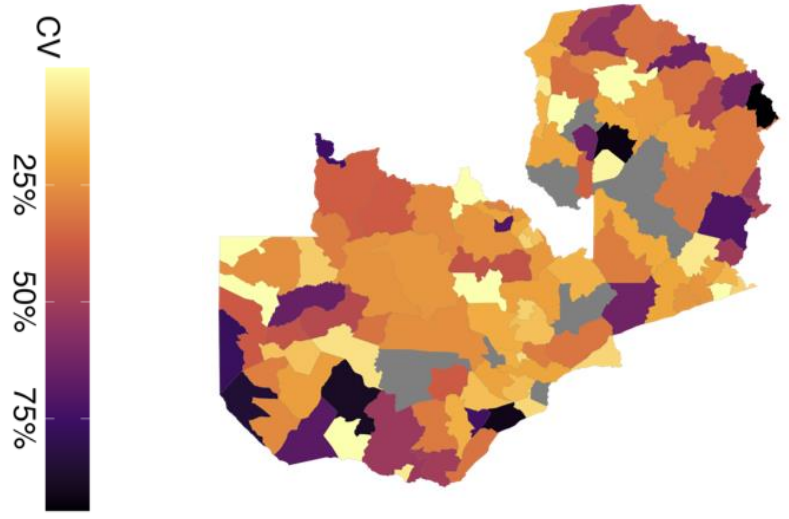


Direct estimates break down









Small Area Estimation (SAE)

SAE = Methods for producing estimates when data is too sparse for subnational direct estimates. I.e., too few/no samples in some areas.

Two main types:

- **Area-level models**: Smooth the direct estimates and their variances to improve accuracy and precision
- **Unit-level models**: Model the outcomes of individual women and aggregate to area-level estimates

Area-level models fully account for the design, but unit-level models need an assumption that the design is ignorable under the given model.

Areal-level models

Assume **observation model** for direct estimate, e.g.,

$$\text{logit}(\hat{P}_a)|p_a, \hat{V}_a \sim N\big(\text{logit}(p_a), \hat{V}_a\big),$$

where $\hat{V}_a$ is estimated variance.

Introduce **latent model** for prevalences

$$\text{logit}(p_a) = \beta_0 + \boldsymbol{x}_a^{\mathrm{T}}\boldsymbol{\beta} + u_a, \qquad a = 1,2,\dots,A,$$

where $\boldsymbol{x}_a$ are area-specific covariates, $\boldsymbol{\beta}$ are coefficients, and u_a are random effects (BYM model).

Estimands are true prevalences $p_1, p_2, \dots, p_A$.

Variance smoothing: we would also model $\hat{V}_1, \hat{V}_2, \dots, \hat{V}_A$. Not done in this talk.

Unit-level models

Overdispersed observation model for cluster c in area a

$$y_{a,c}|r^*_{a,c} \sim \text{Binomial}(n_{a,c}, r^*_{a,c})$$
$$r^*_{a,c}|r_{a,c} \sim \text{Beta}\big(r_{a,c}, d\big),$$

where $y_{a,c}$ is sampled HIV positive, $n_{a,c}$ is sampled women, $r_{a,c}$ is the risk, and d is an overdispersion parameter.

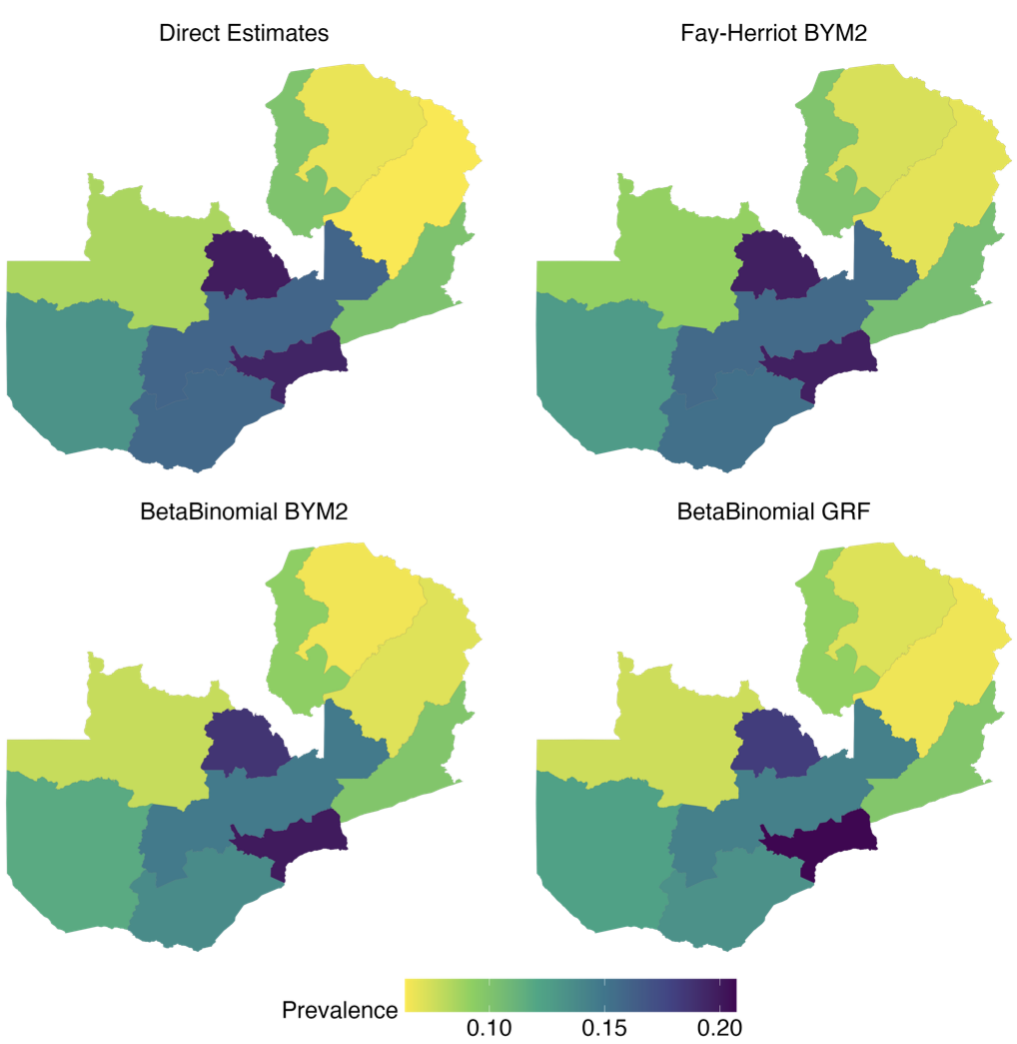
Introduce **latent model** for risks

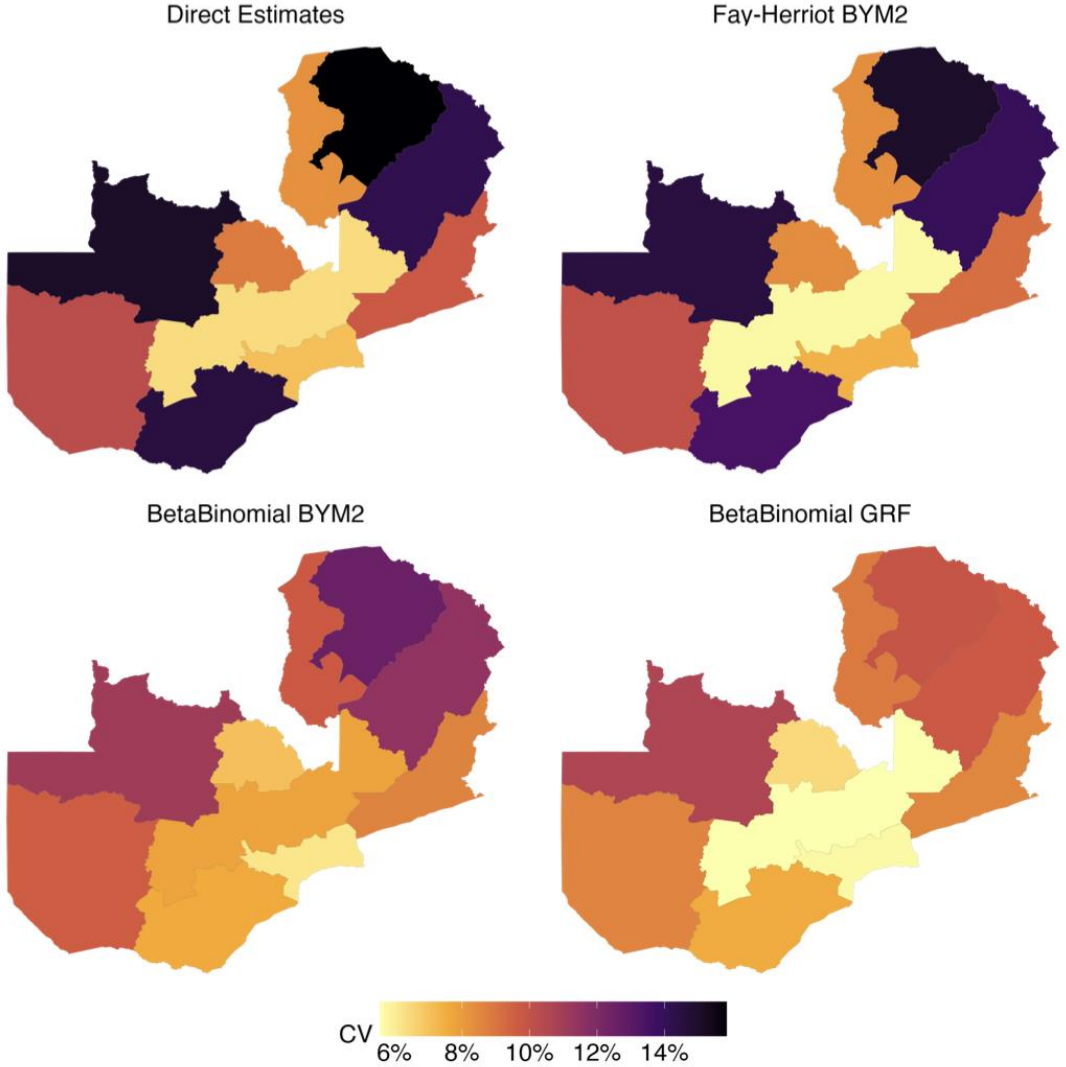
$$\text{logit}(r_{a,c}) = \beta_0 + \boldsymbol{x}^{\text{T}}_{a,c}\boldsymbol{\beta} + u_{a,c},$$

where $\boldsymbol{x}_{a,c}$ are cluster-specific covariates, $\boldsymbol{\beta}$ are coefficients, and $u_{a,c}$ are spatial random effects (BYM2 or GRF).

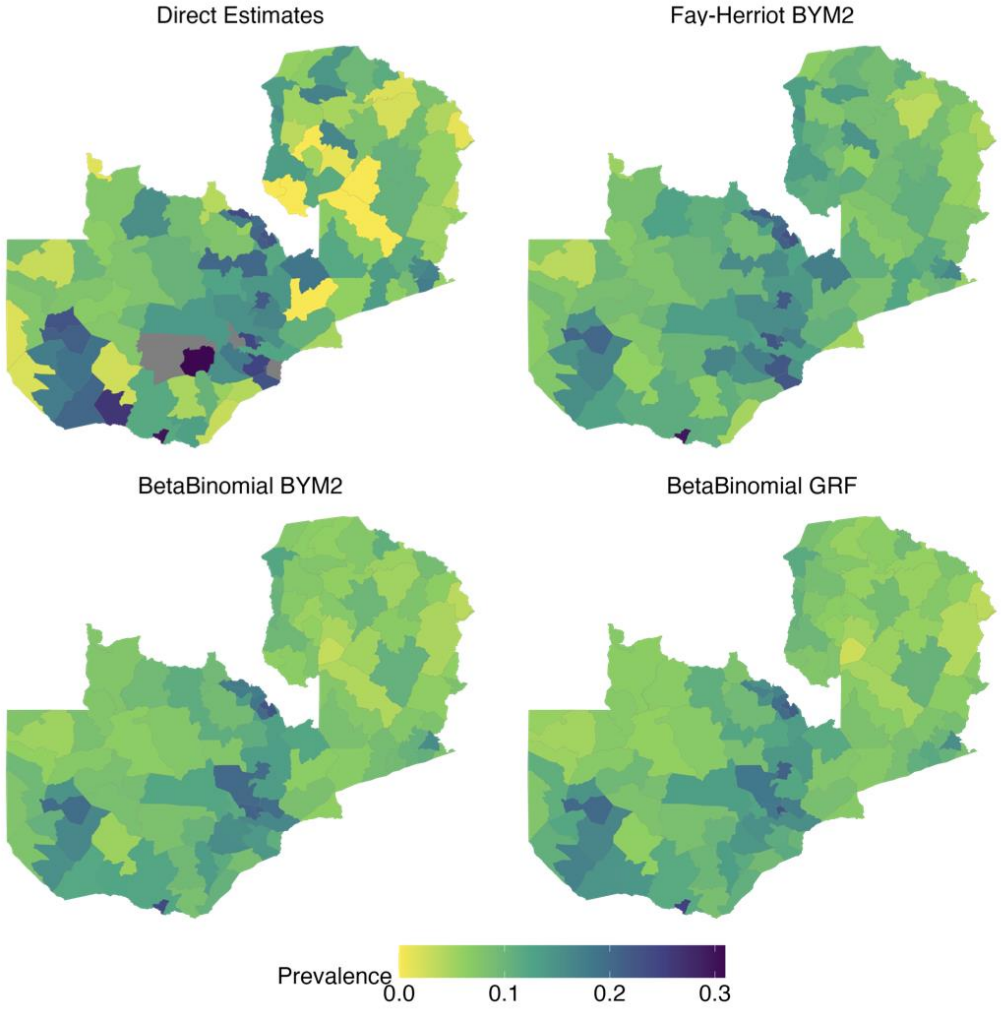
NB: Aggregation to areal estimates is challenging, but no time in this talk.

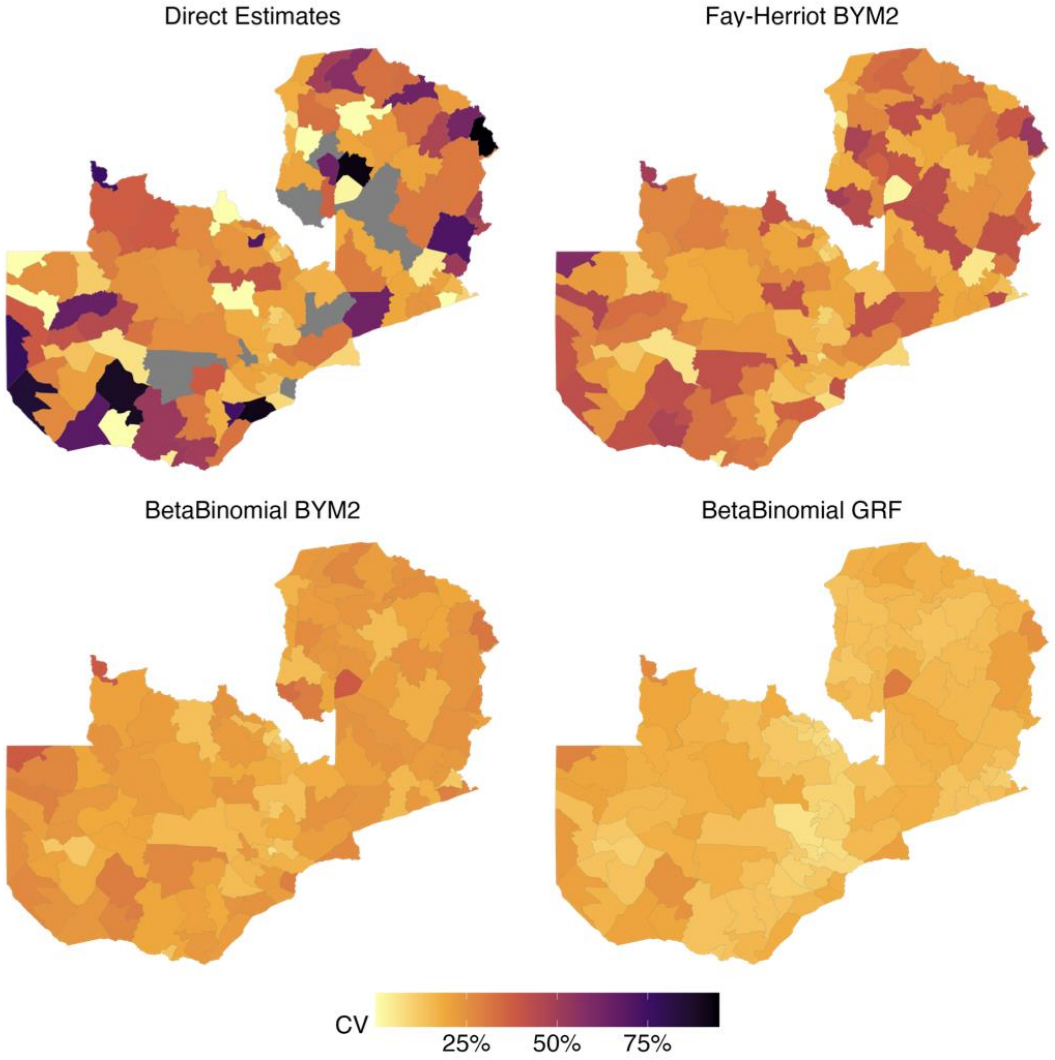
Admin1 – Provinces





Admin2 – Districts





Summary

- Direct estimates are great when they work
- When direct estimates do not work, areal models are nice since they account for the design
- Unit-level models always work, but
  - need to be careful about design
  - give maybe too little uncertainty and too much smoothing
- The choice of model should depend on the data sparsity and target resolution
- Note that estimates from, e.g., IHME arise from models, and are uncertain and sensitive to assumptions and data quality
- Machine learning models are being explored, but data is sparse