Methods and applications of Bayesian spatio-temporal statistics for small-area estimation of HIV indicators

Imperial College London

Adam Howes

Department of Mathematics

Imperial College London

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

September 2023

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of Originality

This thesis, and the work presented in it, is work that I conducted myself. In all cases where I describe others' work, I provide appropriate references.

For someone, or something

Acknowledgements

Thanks to Jeff Eaton and Seth Flaxman for supervision of this research; staff and students of the StatML CDT at Imperial and Oxford; members of the HIV Inference Group at Imperial; members of the Machine Learning and Global Health Network; the Bill & Melinda Gates Foundation and EPSRC for funding this PhD; Mike McLaren, Kevin Esvelt, the Nucleic Acid Observatory team, and the Sculpting Evolution lab for hosting my visit to MIT; Alex Stringer for hosting my visit to Waterloo; the Effective Altruism community; my friends and family.

Adam Howes
Imperial College London
2023

Abstract

Progress towards ending AIDS as a public health threat by 2030 is faltering. Effective public health response requires accurate, timely, high-resolution estimates of epidemic and demographic indicators. Limitations of available data make obtaining such estimates difficult. I develop and apply Bayesian spatio-temporal methods to meet this challenge. Firstly, I examine models for area-level spatial structure. Secondly, I estimate district-level HIV risk group proportions, enabling behavioural prioritisation of prevention services, as put forward by the Global AIDS Strategy. Finally, I develop a novel Bayesian inference method, combining adaptive Gauss-Hermite quadrature with principal component analysis, motivated by the Naomi district-level model of HIV indicators. In sum, the contributions in this thesis help to guide precision HIV policy in sub-Saharan Africa, as well as advancing Bayesian methods for spatio-temporal data.

Contents

List of Figures

List of Tables

List of Abbreviations

HIV . . . . . Human Immunodeficiency Virus.

AIDS . . . . . Acquired ImmunoDeficiency Syndrome.

PEPFAR . . . President's Emergency Plan for AIDS Relief.

HIV . . . . . Demographic and Health Surveys.

AIS . . . . . AIDS Indicator Survey.

PrEP . . . . . Pre-Exposure Prophylaxis.

PEP . . . . . Post-Exposure Prophylaxis.

FSW . . . . . Female Sex Worker(s).

MSM . . . . . Men who have Sex with Men.

PWID . . . . . People Who Inject Drugs.

ANC . . . . . Antenatal Clinic.

UNAIDS . . . United Nations Joint Programme on HIV/AIDS.

CDC . . . . . Centers for Disease Control and Prevention.

UAT . . . . . Unlinked Anonymous Testing.

PMTCT . . . Prevention of Mother-to-Child Transmission.

PLHIV . . . . People Living with HIV.

MCMC . . . . Markov Chain Monte Carlo.

VI . . . . . Variational Inference.

INLA . . . . . Integrated Nested Laplace Approximation.

GP . . . . . Gaussian Process.

CAR . . . . . Conditionally Auto-regressive.

ART . . . . . Antiretroviral Therapy.

SAE . . . . . Small Area Estimation.

GMRF . . . . Gaussian Markov Random Field.

HMC . . . . . Hamiltonian Monte Carlo.

GMRF . . . . Gaussian Markov Random Field.

HMC . . . . . Hamiltonian Monte Carlo.

LGM . . . . . Latent Gaussian Model.

ELGM . . . . . Extended Latent Gaussian Model.

List of Notations

ρ	HIV prevalence.
λ	HIV incidence.
α	ART coverage.
$\mathcal{S}$	Spatial study region $\mathcal{S} \subseteq \mathbb{R}^2$.
$s \in \mathcal{S}$	Point location.
$\mathcal{T}$	Temporal study period $\mathcal{T} \subseteq \mathbb{R}$.
$t \in \mathcal{T}$	Time.

1 Introduction

This thesis is about applied and methodological Bayesian statistics. It is Bayesian in the sense that I use probability models to arrive at conclusions based on data. It is applied and methodological in the sense that I am concerned with real world questions and the means to answer them.

The real world questions relate to surveillance of the HIV epidemic in sub-Saharan Africa. Though progress has been made, millions of people are impacted by HIV each year. Quantifying the epidemic using statistics is an important part of the public health response, and the path towards disease control and elimination.

The data in this thesis relate to people answering survey questions or using healthcare facilities. As such, it has particular positions in space and time which should be taken into account. Spatio-temporal data, while encompassing a great diversity, has distinctive commonalities which make its collective study worthwhile.

Computation is an essential part of modern statistical practice. Each project chapter is accompanied by code, hosted on GitHub.

1.1 Chapter overview

- Chapter 2: I start by reviewing the required background for the rest of the thesis, namely regarding the HIV/AIDS epidemic and Bayesian spatio-

temporal statistics.

- Chapter 3: The prevailing model for spatial structure used in small-area estimation (Besag et al. 1991) was designed with analysis of a grid of pixels in mind. In disease mapping, we work using the districts of a country, which are not a grid. I evaluate the practical consequences of this this concern (Howes, Eaton, et al. 2023+).
- Chapter 4: Adolescent girls and young women are a demographic group at disproportionate risk of HIV infection. The Global AIDS Strategy suggests prioritising interventions on the basis of behaviour to prevent the most new infections using available resources. I estimate the size of behavioural risk groups across priority countries to enable implementation of this strategy, and assess the potential benefits in terms of numbers of new infections prevented (Howes, Risher, et al. 2023).
- Chapter 5: The Naomi small-area estimation model (Eaton et al. 2021) is used by countries to estimate district-level HIV indicators. With this motivation, I develop an approximate Bayesian inference method combining adaptive Gauss-Hermite quadrature with principal components analysis (Howes, Stringer, et al. 2023+). I apply the method to data from Malawi, and analyse the consequence of inference method choice for policy relevant outcomes. Further, I open the door to a new class of fast, flexible, and accurate Bayesian inference algorithms.
- Chapter 6: I discuss avenues for future work, and my conclusions regarding the research.

2
Background

2.1 The HIV/AIDS epidemic

Human immunodeficiency virus (HIV) is a retrovirus which infects humans. HIV primarily infects CD4+ T cells, a type of white blood cell vital for the function of the immune system. Over time, if left untreated HIV can progress to a more advanced stage known as acquired immunodeficiency syndrome (AIDS). AIDS is characterised by increased risk of developing infections normally controlled by the immune system.

HIV is transmitted by exposure to certain bodily fluids of an infected person. The most common mode of transmission is through sexual contact.

The first HIV cases were reported in the early 1980s. In 2021 there were thirty-eight million people living with HIV, six hundred fifty thousand AIDS-related deaths, and one million, five hundred thousand people newly infected with HIV (UNAIDS 2021b).

A major global effort has been made to address the epidemic. Significant progress has been made, both in reducing the number of new HIV cases and decreasing the number of AIDS related deaths (Figure 2.1). Roll out of antiretroviral therapy (ART) has been a key tool. Treatment Other interventions include condoms, pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP), and voluntary

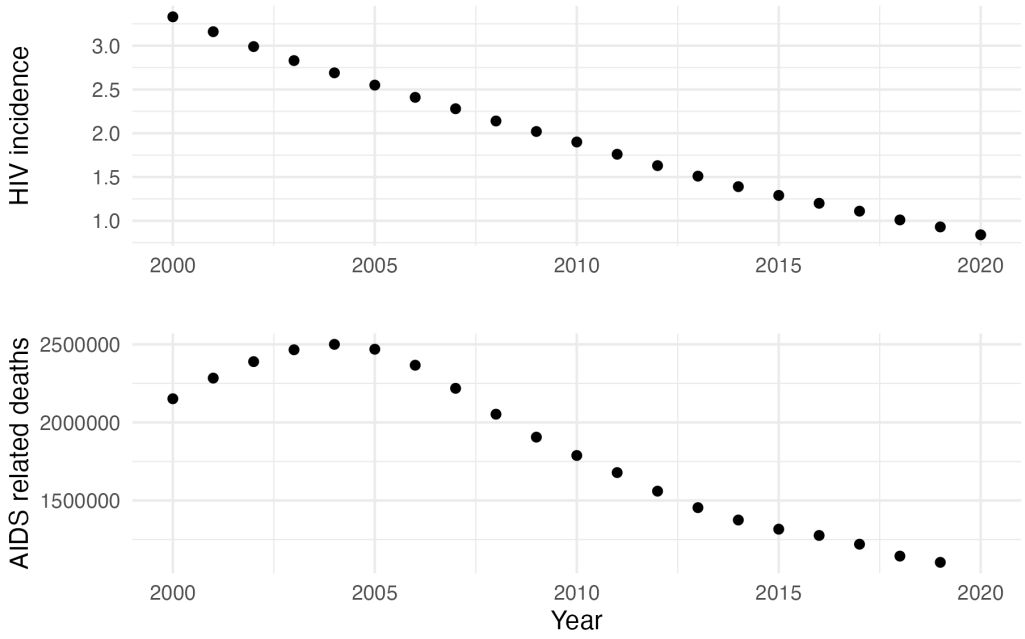

Figure 2.1: Overall picture.

medical male circumcision (VMMC).

There is substantial geographic inequality in disease burden. In some countries, the epidemic is concentrated within small populations, and prevalence is low. For others, transmission is sustained in the general population, and prevalence is higher. Most of the countries severely affected are in sub-Saharan Africa (SSA), accounting for percentage of people living with HIV (PLHIV) worldwide. There is also significant geographic variation within countries.

In all settings, there is substantial inequality in disease burden across groups. Groups at increased risk of HIV infection are known as key populations (KPs), and include men who have sex with men (MSM), female sex workers (FSW), people who inject drugs (PWID), and transgender people (TGP). KPs are often marginalised, and face legal and social issues.

In SSA, risk is more diffuse across the population than in concentrated settings. Large demographic groups at increased risk of HIV infection include adolescent girls and young women (AGYW).

HIV interventions should be prioritised to have the greatest impact on the

epidemic. The precision public health paradigm aims to get the right interventions, to the right populations, in the right place, at the right time.

HIV surveillance refers to the collection, analysis, interpretation and dissemination of data relating to HIV/AIDS. Surveillance can used to track epidemic indicators, identify at-risk populations, find drivers of transmission, and evaluate the impact of prevention and treatment programs. Important indicators include HIV prevalence, which is the proportion of the population who have HIV; HIV incidence, the rate of new HIV infections; and ART coverage, which is the proportion of PLHIV who are on ART.

There are significant difficulties associated with furnishing these estimates. The difficulties include data sparsity in space and time, survey bias, conflicting information sources, hard to reach populations, and changing demographics. These data limitations foreground the importance of statistical modelling, including synthesising multiple sources of information.

Aims for HIV response going forward, and surveillance capabilities are needed to meet them. Phasing out of nationally-representative household surveys for HIV.

2.2 Bayesian spatio-temporal statistics

2.2.1 Bayesian statistics

Bayesian statistics is a mathematical paradigm for learning from data. I provide a brief, opinionated, overview in this section, and recommend McElreath (2020) or Gelman et al. (2013) for a more complete introduction.

Bayesian modelling

At its best, the Bayesian paradigm allows the analyst focus their attention on the question of how to model the data. This is achieved by the construction of a generative model $p(\mathbf{y}, \boldsymbol{\vartheta})$ for the observed data $\mathbf{y}$ together with parameters $\boldsymbol{\vartheta}$. The model is generative in the sense that one can simulate from it to obtain draws $(\mathbf{y}, \boldsymbol{\vartheta}) \sim p(\mathbf{y}, \boldsymbol{\vartheta})$. If these draws differ too greatly from what the analyst

would expect, then the generative model can be refined. This is what is known as a prior predictive check.

The model is usually constructed from two parts, known as the likelihood $p(\mathbf{y} \mid \boldsymbol{\vartheta})$ and the prior $p(\boldsymbol{\vartheta})$ such that $p(\mathbf{y}, \boldsymbol{\vartheta}) = p(\mathbf{y} \mid \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})$. The likelihood, as a function of $\boldsymbol{\vartheta}$ with $\mathbf{y}$ fixed, reflects the probability of observing the data when the true value of the parameters is $\boldsymbol{\vartheta}$. The prior encapsulates beliefs about the parameters $\boldsymbol{\vartheta}$ before the data is observed.

There is disagreement about how the prior should be specified. The distinction between likelihood and prior can sometimes be blurred (Section 2.2.3).

Bayesian computation

Interest lies in obtaining the posterior distribution $p(\boldsymbol{\vartheta} \mid \mathbf{y})$ which represents beliefs about the parameters given the observed data. Using Bayes' theorem, the posterior distribution is given by

$$p(\boldsymbol{\vartheta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})}{p(\mathbf{y})}. \tag{2.1}$$

However, it is usually intractable to calculate the posterior distribution directly because of the integral $p(\mathbf{y}) = \int p(\mathbf{y}, \boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta}$ in the denominator. As such, though the numerator is proportional to the posterior $p(\boldsymbol{\vartheta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})$ and easy to evaluate, it is not easy to evaluate the posterior itself. A great variety of computational methods have been developed at to tackle this problem. Markov chain Monte Carlo (MCMC) is the most popular approach, and proceeds by simulating samples from a Markov chain with the posterior as its stationary distribution. Variational Bayes approaches assume the posterior distribution belongs to a certain class of functions and use optimisation to choose the best member of that class.

Interplay between modelling and computation

Bayesian computation aspires to abstract away calculation of the posterior distribution. Modern computational techniques and software have made this aspiration a reality for many models. However, computation of the posterior

remains intractable for a substantial majority of models. As such, the analyst need not only to be concerned with choosing a model suitable for the data, but also choosing a model for which the posterior is tractable in reasonable time. It is in this sense, that there is an interplay between modelling and computation. As computation improves, the space of models available to the analyst expands.

2.2.2 Spatio-temporal statistics

In spatio-temporal statistics (Cressie and Wikle 2015) we observe data indexed by spatial or temporal location. In this thesis we assume that the spatial study region $\mathcal{S} \subseteq \mathbb{R}^2$ has two dimensions, corresponding to latitude and longitude. Data may be associated to a point $s \in \mathcal{S}$ or area $A \subseteq \mathcal{S}$ in the study region. The temporal study period $\mathcal{T} \subseteq \mathbb{R}$ can more generally be assumed to be one dimensional.

Commonly used independent and identically distributed (IID) assumptions on observations are rarely suitable in this setting because we expect there to be spatio-temporal correlation structure.

2.2.3 Model classes

Hierarchical models

Bayesian hierarchical models are comprised of multiple stages

$$p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Latent Gaussian models

Latent Gaussian models [LGMs; Rue et al. (2009)] are a class of three-stage Bayesian hierarchical models in which, loosely speaking, the middle layer is Gaussian. More specifically, in an LGM, the likelihood is given by

$$\begin{aligned} y_i &\sim p(y_i \,|\, \eta_i, \theta_1), \quad i \in [n] \\ \mu_i &= \mathbb{E}(y_i \,|\, \eta_i) = g(\eta_i), \\ \eta_i &= \beta_0 + \sum_{l=1}^{p} \beta_j z_{ji} + \sum_{k=1}^{r} f_k(u_{ki}), \end{aligned}$$

where $[n] = \{1, \ldots, n\}$. The response variable is $\mathbf{y} = (y)_{i \in [n]}$ with likelihood $p(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}_1) = \prod_{i=1}^{n} p(y_i \mid \eta_i, \boldsymbol{\theta}_1)$, where $\boldsymbol{\eta} = (\eta)_{i \in [n]}$. Each response has conditional mean μ_i with inverse link function $g : \mathbb{R} \to \mathbb{R}$ such that $\mu_i = g(\eta_i)$. The vector $\boldsymbol{\theta}_1 \in \mathbb{R}^{s_1}$, with s_1 assumed small, are additional parameters of the likelihood. The structured additive predictor η_i may include an intercept β_0, linear effects β_j of the covariates z_{ji}, and unknown functions $f_k(\cdot)$ of the covariates u_{ki}. The parameters β_0, $\{\beta_j\}$, $\{f_k(\cdot)\}$ are each assigned Gaussian priors, and can be collected into a vector $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}_2)^{-1})$ where $\boldsymbol{\theta}_2 \in \mathbb{R}^{s_2}$ are further parameters, again with s_2 assumed small. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^m$ with $m = s_1 + s_2$ be all hyperparameters, with prior $p(\theta)$. In total, the parameters of the LGM $\boldsymbol{\vartheta} = (\mathbf{x}, \boldsymbol{\theta})$ comprise both the latent field and hyperparameters.

Spatio-temporal data are well suited to being modelled with LGMs.

Extended latent Gaussian models

Many of leading-edge disease mapping models fall outside the LGM class. However, many of these models do fit into the class of extended latent Gaussian models [ELGMs; Stringer et al. (2021)]. By allowing many-to-one link functions, ELGMs facilitate modelling of non-linearities. The structured additive predictor is redefined as $\boldsymbol{\eta} = (\eta)_{i \in [N_n]}$, where $N_n \in \mathbb{N}$ is a function of n, and it is possible that $N_n \neq n$. Each mean response μ_i now depends on some subset $\mathcal{J}_i \subseteq [N_n]$ of indices of $\boldsymbol{\eta}$, with $\cup_{i=1}^{n} \mathcal{J}_i = [N_n]$ and $1 \leq |\mathcal{J}_i| \leq N_n$. The inverse link function $g(\cdot)$ is redefined for each observation to be a possibly many-to-one mapping $g_i : \mathbb{R}^{|\mathcal{J}_i|} \to \mathbb{R}$, such that $\mu_i = g_i(\boldsymbol{\eta}_{\mathcal{J}_i})$. Put together, ELGMs are then of the form

$$
\begin{aligned}
y_i &\sim p(y_i \mid \boldsymbol{\eta}_{\mathcal{J}_i}, \boldsymbol{\theta}_1), \quad i \in [n] \\
\mu_i &= \mathbb{E}(y_i \mid \boldsymbol{\eta}_{\mathcal{J}_i}) = g_i(\boldsymbol{\eta}_{\mathcal{J}_i}), \\
\eta_j &= \beta_0 + \sum_{l=1}^{p} \beta_j z_{ji} + \sum_{k=1}^{r} f_k(u_{ki}), \quad j \in [N_n],
\end{aligned}
$$

with latent field and hyperparameter priors as in the LGM case.

3

Spatial structure

In this chapter, I describe an investigation of spatial random effects specifications. The motivated for this investigation was one of the fundamental questions encountered by an analyst during model construction. Namely, should the model be augmented to better capture a feature of the data generating process that we believe exists? The results are presented in Howes, Eaton, et al. (2023+).

Code for the analysis in this chapter is available from `athowes/beyond-borders` and supported by the R package `arealutils`.

3.1 Background

3.1.1 Areal and point data

3.1.2 Spatial random effects

3.2 Models based on adjacency

3.2.1 The Besag model

3.2.2 The BYM2 model

3.3 Models using kernels

3.3.1 The centroid kernel model

3.3.2 The integrated kernel model

3.4 Simulation study

3.4.1 Synthetic data-sets

3.4.2 Inferential models

Priors

Kernel details

3.4.3 Inference algorithms

3.4.4 Model assessment

Continuous ranked probability score

3.4.5 Results

3.5 HIV prevalence study

3.5.1 Results

3.6 Discussion

3.6.1 Limitations

3.6.2 Conclusion

4

A model for risk group proportions

In this chapter I describe an application of Bayesian spatio-temporal statistics to small-area estimation of HIV risk group proportions. This work was conducted in collaboration with colleagues from the MRC Centre for Global Infectious Disease Analysis and UNAIDS. My primary role was to develop the statistical model. I built on an earlier version of the analysis conducted by Kathryn Risher. The results are presented in Howes, Risher, et al. (2023). Kathryn has also created a spreadsheet tool using the estimates which is now being used by countries to guide policy. Code for the analysis in this chapter is available from `athowes/multi-agyw` and supported by the R package `multi.utils` (Howes 2022).

4.1 Background

Adolescent girls and young women (AGYW, here defined as females aged 15-29) are a demographic group at increased risk of HIV infection AGYW comprise 44% of new infections, while only 28% of the population (UNAIDS 2021a). HIV incidence for AGYW is 2.4 times higher than for similarly aged males. The reasons for this disparity are given in Table 4.1.

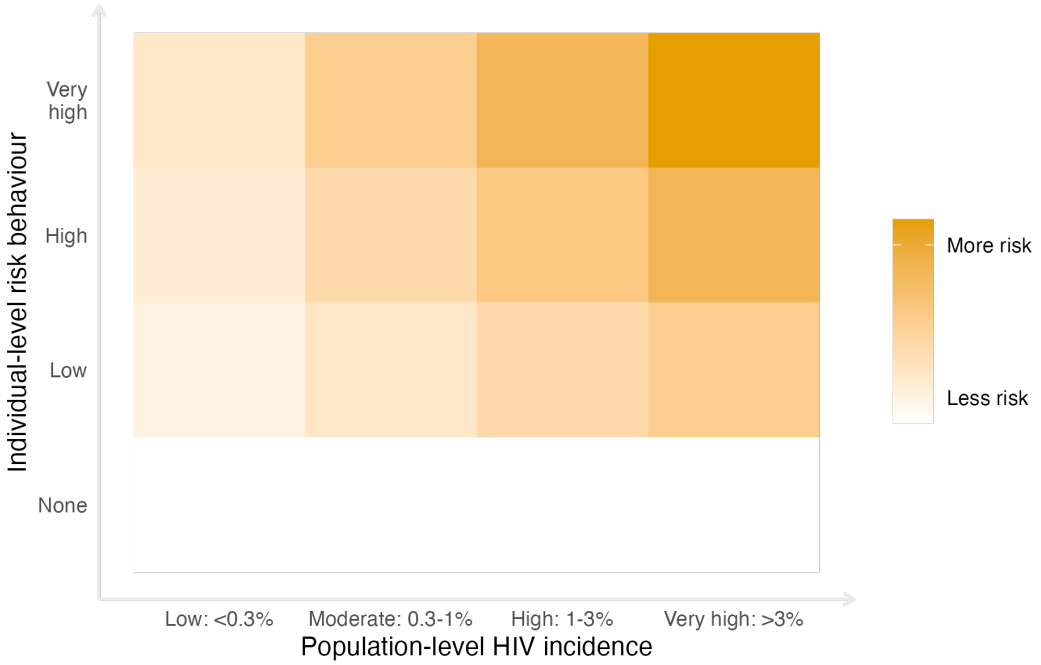

Figure 4.1: Risk depends on both individual-level risk behaviour and population-level HIV incidence.

Table 4.1: Reasons.

Reason	Description
Structural vulnerability and power imbalances	Text
Age patterns of sexual mixing	Text
Younger age at first sex	Text
Increased susceptibility to HIV infection	Text

On this basis, AGYW have been identified as a priority population for HIV prevention services (Saul et al. 2018; The Global Fund 2018). The Global AIDS Strategy 2021-2026 (UNAIDS 2021b) proposed stratifying HIV prevention packages to AGYW based on 1) local population-level HIV incidence and 2) individual-level sexual risk behaviour. As risk depends on both factors, prioritisation of prevention services would be more efficient if both are taken into account (Figure 4.1). The strategy encourages programmes to define targets for the proportion of AGYW to be reached with a range of interventions. Estimates of the size of each risk group are required.

Table 4.2: Behavioural risk groups.

Risk group	Description	Local HIV incidence	Incidence ratio
None	Not sexually active	–	0.0
Low	One cohabiting partner	–	1.0 (Baseline)
High	Non-regular or multiple partner(s)	–	1.72
Very High	Transactional sex (adjusted to correspond to female sex workers)	0.1-0.3%	13.0
		0.3-1.0%	9.0
		1.0-3.0%	6.0
		>3.0%	3.0

4.2 Data

I used household survey data from 13 countries: Botswana, Cameroon, Kenya, Lesotho, Malawi, Mozambique, Namibia, South Africa, Eswatini, Tanzania, Uganda, Zambia and Zimbabwe. These countries have been designated AGYW priority countries.

For each survey, I classified respondents into one of four behavioural risk groups according to reported sexual risk behaviour in the past 12 months. These risk groups were: not sexually active, one cohabiting sexual partner, non-regular or multiple sexual partner(s), and AGYW who report transactional sex. In the case of inconsistent responses, women were categorised according to the highest risk group they fell into, ensuring that the categories were mutually exclusive. Exact survey questions varied slightly across survey types and between survey phases. Questions captured information about whether the respondent had been sexually active in the past twelve months, and if so how with many partners. For their three most recent partners, respondents were also asked about the type of partnership

(spouse, cohabiting partner, partner not cohabiting with respondent, friend, sex worker, sex work client, and other).

Some surveys included a specific question asking if the respondent had received or given money or gifts for sex in the past twelve months. In these surveys, 2.64% of women reported transactional sex. In surveys without such a question, women almost never (0.01%) answered that one of their three most recent partners was a sex work client. Due to this incomparability across surveys, I did not include surveys without a specific transactional sex question when estimating the proportion of the population who engaged in transactional sex. I instead focused on estimating the proportion of women who reported transactional sex at a district level, and subsequently adjusted these proportions to align to national estimates for the number of female sex workers.

I used estimates of population, PLHIV and new HIV infections stratified by district and age group from HIV estimates published by UNAIDS that were developed using the Naomi model (Eaton et al. 2021). The administrative area hierarchy and geographic boundaries I used correspond to those used for health service planning by countries, exceptions being Cameroon and Kenya where I conducted analysis one level higher at the department and county levels, respectively. I used the most recent 2022 estimates for all countries, apart from Mozambique where, due to data accuracy concerns, I used the 2021 estimates (in which the Cabo Delgado province is excluded due to disruption by conflict).

4.3 Model for risk group proportions

I took a two-stage modelling approach to estimate the four risk group proportions. Index the four risk groups as $k \in \{1, 2, 3, 4\}$, and denote being in either the third or fourth risk group by $k = 3^+$. First, using all the surveys, I used a spatio-temporal multinomial logistic regression model (Section 4.3.1) to estimate the proportion of AGYW in the risk groups $k \in \{1, 2, 3^+\}$. Then, using only those surveys with a specific transactional sex question, I fit a spatial logistic regression model (Section

4.3.2) to estimate the proportion of those in the $k = 3^+$ risk group that were in the $k = 3$ and $k = 4$ risk groups respectively.

4.3.1 Spatio-temporal multinomial logistic regression

Let $i \in \{1, \ldots, n\}$ denote districts partitioning the 13 studied AGYW priority countries $c[i] \in \{1, \ldots, 13\}$. Consider the years 1999-2018 denoted as $t \in \{1, \ldots, T\}$, and age groups $a \in \{15\text{–}19, 20\text{–}24, 25\text{–}29\}$. Let $p_{itak} > 0$ with $\sum_{k=1}^{3^+} p_{itak} = 1$, be the probabilities of membership of risk group k.

Multinomial logistic regression

A baseline category multinomial logistic regression model is specified by

$$\mathbf{y}_{ita} = (y_{ita1}, \ldots, y_{ita3^+})^\top \sim \text{Multinomial}(m_{ita};\, p_{ita1}, \ldots, p_{ita3^+}), \tag{4.1}$$

$$\log\left(\frac{p_{itak}}{p_{ita1}}\right) = \eta_{itak}, \quad k = 2, 3^+, \tag{4.2}$$

where the number in risk group k is y_{itak}, the fixed sample size is $m_{ita} = \sum_{k=1}^{3^+} y_{itak}$, and $k = 1$ is chosen as the baseline category. This model is not an LGM, and is not possible to fit in `R-INLA`.

The multinomial-Poisson transformation

I use the multinomial-Poisson transformation to enable inference with `R-INLA`. The transformation reframes a given multinomial logisitic regression model as an equivalent Poisson log-linear model of the form

$$y_{itak} \sim \text{Poisson}(\kappa_{itak}), \tag{4.3}$$

$$\log(\kappa_{itak}) = \eta_{itak}. \tag{4.4}$$

The basis of the transformation is that, conditional on their sum, Poisson counts are jointly multinomially distributed (McCullagh and Nelder 1989) as follows

$$\mathbf{y}_{ita} \,|\, m_{ita} \sim \text{Multinomial}\left(m_{ita};\, \frac{\kappa_{ita1}}{\kappa_{ita}}, \ldots, \frac{\kappa_{ita3^+}}{\kappa_{ita}}\right), \tag{4.5}$$

where $\kappa_{ita} = \sum_{k=1}^{3^+} \kappa_{itak}$. Category probabilities are then obtained by the softmax function

$$p_{itak} = \frac{\exp(\eta_{itak})}{\sum_{k=1}^{3^+} \exp(\eta_{itak})} = \frac{\kappa_{itak}}{\sum_{k=1}^{3^+} \kappa_{itak}} = \frac{\kappa_{itak}}{\kappa_{ita}}. \tag{4.6}$$

Under the equivalent model, the sample sizes $m_{ita} = \sum_k y_{itak}$ are treated as random $m_{ita} \sim \text{Poisson}\,(\kappa_{ita})$ rather than fixed. The joint distribution of $p(\mathbf{y}_{ita}, m_{ita}) = p(\mathbf{y}_{ita} \,|\, m_{ita})p(m_{ita})$ is then

$$p(\mathbf{y}_{ita}, m_{ita}) = \exp(-\kappa_{ita})\frac{(\kappa_{ita})^{m_{ita}}}{m_{ita}!} \times \frac{m_{ita}!}{\prod_k y_{itak}!} \prod_k \left(\frac{\kappa_{itak}}{\kappa_{ita}}\right)^{y_{itak}} \tag{4.7}$$

$$= \prod_k \left(\frac{\exp(-\kappa_{itak})\,(\kappa_{itak})^{y_{itak}}}{y_{itak}!}\right) \tag{4.8}$$

$$= \prod_k \text{Poisson}\,(y_{itak} \,|\, \kappa_{itak})\,. \tag{4.9}$$

corresponding to the product of independent Poisson likelihoods as in Equation 4.3.

This model, including random sample sizes, is equivalent to the multinomial logistic regression only when the normalisation constants m_{ita} are recovered exactly. To ensure that this is the case, one approach is to include observation-specific random effects θ_{ita} in the equation for the linear predictor. Multiplying each of $\{\kappa_{itak}\}_{k=1}^{3^+}$ by $\exp(\theta_{ita})$ has no effect on the category probabilities, but does provide the necessary flexibility for κ_{ita} to recover m_{ita} exactly. Although in theory an improper prior $\theta_{ita} \propto 1$ should be used, I found that in practise, by keeping η_{ita} otherwise small using appropriate constraints, so that arbitrarily large values of θ_{ita} are not required, it is sufficient (and practically preferable for inference) to instead use a vague prior.

Model specifications

I considered four models for η_{ita} in Equation 4.4 of the form

$$\eta_{ita} = \theta_{ita} + \beta_k + \zeta_{c[i]k} + \alpha_{ac[i]k} + \phi_{ik} + \gamma_{tk}. \tag{4.10}$$

Observation random effects $\theta_{ita} \sim \mathcal{N}(0, 1000^2)$ were included in all models I considered, and are required for the multinomial-Poisson transformation to be valid. To capture country-specific proportion estimates for each category, I included category random effects $\beta_k \sim \mathcal{N}(0, \tau_\beta^{-1})$ and country-category random effects

$\zeta_{ck} \sim \mathcal{N}(0, \tau_\zeta^{-1})$. Heterogeneity in risk group proportions by age was allowed by including age-country-category random effects $\alpha_{ack} \sim \mathcal{N}(0, \tau_\alpha^{-1})$.

Spatial random effects For the space-category ϕ_{ik} random effects I considered two specifications:

1. Independent and identically distributed (IID) $\phi_{ik} \sim \mathcal{N}(0, \tau_\phi^{-1})$,
2. Besag (Besag et al. 1991) grouped by category

$$\boldsymbol{\phi} = (\phi_{11}, \ldots, \phi_{n1}, \ldots, \phi_{13^+}, \ldots \phi_{n3^+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_\phi \mathbf{R}_\phi^\star)^-).$$

   The scaled structure matrix $\mathbf{R}_\phi^\star = \mathbf{R}_b^\star \otimes \mathbf{I}$ is given by the Kronecker product of the scaled Besag structure matrix $\mathbf{R}_b^\star$ and the identity matrix $\mathbf{I}$, and $-$ denotes the generalised matrix inverse. Scaling of the structure matrix to have generalised variance one ensures interpretable priors may be placed on the precision parameter (Sørbye and Rue 2014). I followed the further recommendations of Freni-Sterrantino et al. (2018) with regard to disconnected adjacency graphs, singletons and constraints. The Besag structure matrix $\mathbf{R}_b$ was obtained by the precision matrix of the random effects $\mathbf{b} = (b_1, \ldots, b_n)^\top$ with full conditionals

$$b_i \,|\, \mathbf{b}_{-i} \sim \mathcal{N}\left(\frac{\sum_{j:j\sim i} b_j}{n_{\delta i}}, \frac{1}{n_{\delta i}}\right), \tag{4.11}$$

   where $j \sim i$ if the districts A_i and A_j are adjacent, and $n_{\delta i}$ is the number of districts adjacent to A_i.

In preliminary testing, I excluded spatial random effects from the model, but found that this negatively effected performance. I also tested using the BYM2 model (Simpson et al. 2017) in place of the Besag, but found that the proportion parameter posteriors tended to be highly peaked at the value one. For simplicity and to avoid numerical issues, by using Besag random effects I effectively decided to fix this proportion to one.

Temporal random effects For the year-category γ_{tk} random effects I considered two specifications:

1. IID $\phi_{tk} \sim \mathcal{N}(0, \tau_\phi^{-1})$,
2. First order autoregressive (AR1) grouped by category

$$\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{13^+}, \ldots, \gamma_{T1}, \ldots, \gamma_{T3^+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_\phi \mathbf{R}_\gamma^\star)^-).$$

The scaled structure matrix $\mathbf{R}_\gamma^\star = \mathbf{R}_r^\star \otimes \mathbf{I}$ is given by the Kronecker product of a scaled AR1 structure matrix $\mathbf{R}_r^\star$ and the identity matrix $\mathbf{I}$. The AR1 structure matrix $\mathbf{R}_r$ is obtained by precision matrix of the random effects $\mathbf{r} = (r_1, \ldots, r_T)^\top$ specified by

$$r_1 \sim \left(0, \frac{1}{1-\rho^2}\right), \tag{4.12}$$

$$r_t = \rho r_{t-1} + \epsilon_t, \quad t = 2, \ldots, T, \tag{4.13}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$ and $|\rho| < 1$.

Priors All random effect precision parameters $\tau \in \{\tau_\beta, \tau_\zeta, \tau_\alpha, \tau_\phi, \tau_\gamma\}$ were given independent penalised complexity (PC) priors (Simpson et al. 2017) with base model $\sigma = 0$ given by $p(\tau) = 0.5\nu\tau^{-3/2} \exp\left(-\nu\tau^{-1/2}\right)$ where $\nu = -\ln(0.01)/2.5$ such that $\mathbb{P}(\sigma > 2.5) = 0.01$. For the lag-one correlation parameter ρ, I used the PC prior, as derived by Sørbye and Rue (2017), with base model $\rho = 1$ and condition $\mathbb{P}(\rho > 0 = 0.75)$. I chose the base model $\rho = 1$ corresponding to no change in behaviour over time, rather than the alternative $\rho = 0$ corresponding to no correlation in behaviour over time, as I judged the former to be more plausible a priori.

Constraints

To ensure interpretable posterior inferences relating to the random effects, I applied sum-to-zero constraints such that none of the category interaction random effects altered overall category probabilities. For the space-year-category random effects, I applied analogous sum-to-zero constraints to maintain roles of the space-category and year-category random effects. Together, these were:

A model for risk group proportions

1. Category $\sum_k \beta_k = 0$,
2. Country $\sum_c \zeta_{ck} = 0$, $\forall\, k$,
3. Age-country $\sum_a \alpha_{ack} = 0$, $\forall\, c, k$,
4. Spatial $\sum_i \phi_{ik} = 0$, $\forall\, k$,
5. Temporal $\sum_t \gamma_{tk} = 0$, $\forall\, k$.

Survey weighted likelihood

I included surveys which use a complex design, in which each individual has an unequal probability of being included in the sample. For example the DHS often employs a two-stage cluster design, first taking an urban rural stratified sample of ennumeration areas, before selecting households from each ennumeration area using systematic sampling (DHS 2012).

To account for this aspect of survey design, I use a weighted pseudo-likelihood where the observed counts y are replaced by effective counts $y^\star$ calculated using the survey weights w_j of all individuals j in the corresponding strata. I multiplied direct estimates produced using the `survey` package (Lumley 2004) by the Kish effective sample size (Kish 1965)

$$m^\star = \frac{\left(\sum_j w_j\right)^2}{\sum_j {w_j}^2} \tag{4.14}$$

to obtain $y^\star$. These counts may not be integers, and as such the Poisson likelihood I used in Equation 4.3 is not appropriate. Instead, I used a generalised Poisson pseudo-likelihood $y^\star \sim \text{xPoisson}(\kappa)$, given by

$$p(y^\star) = \frac{\kappa^{y^\star}}{\lfloor y^\star ! \rfloor} \exp\left(-\kappa\right), \tag{4.15}$$

as implemented by `family = "xPoisson"` in `R-INLA`, which accepts non-integer input.

Model selection

4.3.2 Spatial logistic regression

To estimate the proportion of those in the $k = 3^+$ risk group that were in the $k = 3$ and $k = 4$ risk groups respectively, I fit logistic regression models of the form

$$y_{ia4} \sim \text{Binomial}\left(y_{ia3} + y_{ia4}, q_{ia}\right), \tag{4.16}$$

$$q_{ia} = \text{logit}^{-1}\left(\eta_{ia}\right), \tag{4.17}$$

where $q_{ia} = p_{ia4}/(p_{ia3} + p_{ia4}) = p_{ia4}/p_{ia3^+}$. Taking this two-step approach allowed me to include all surveys in the multinomial regression model, but only those surveys with a specific transactional sex question in Equation 4.16. As all such surveys occurred in 2013-2018, in the logistic regression model I assumed q_{ia} to be constant over time.

I considered six logistic regression models each including a constant $\beta_0 \sim \mathcal{N}(-2, 1^2)$, country random effects $\zeta_c \sim \mathcal{N}(0, \tau_\zeta^{-1})$, and age-country random effects $\alpha_{ac} \sim \mathcal{N}(0, \tau_\alpha^{-1})$. The prior on β_0 placed 95% prior probability on the range 2-50% for the percentage of those with non-regular or multiple partners who report transactional sex. I considered two specifications (IID, Besag) for the spatial random effects ϕ_i. To aid estimation with sparse data, I also considered national-level covariates for the proportion of men who have paid for sex ever `cfswever` or in the last twelve months `cfswrecent`, available from Hodgins et al. (2022). For both random effect precision parameters $\tau \in \{\tau_\alpha, \tau_\zeta\}$ I used the PC prior with base model $\sigma = 0$ and $\mathbb{P}(\sigma > 2.5 = 0.01)$. For the regression parameters $\beta \in \{\beta_{\texttt{cfswever}}, \beta_{\texttt{cfswrecent}}\}$ I used the prior $\beta \sim \mathcal{N}(0, 2.5^2)$.

Model specifications
Survey weighted likelihood
Model selection

4.3.3 Coverage assessment

4.3.4 Female sex worker population size adjustment

Responding "yes" to the survey question "have you had sex in return for gifts, cash or anything else in the past 12 months" is not considered sufficient to constitute

sex work. In recognition of this, I adjusted the estimates obtained based on the survey to match FSW population size estimates obtained via alternative methods.

Stevens et al. (2022) used a Bayesian meta-analysis of key population specific data sources to estimate adult (15-49) FSW population size by country. I disaggregated these estimates by age as follows. First, I calculated the total sexually debuted population in each age group, in each country. To describe the distribution of age at first sex, I used skew logistic distributions (Nguyen and Eaton 2022) with cumulative distribution function given by

$$F(x) = (1 + \exp(\kappa_c(\mu_c - x)))^{-\gamma_c}, \tag{4.18}$$

where $\kappa_c, \mu_c, \gamma_c > 0$ are country-specific shape, shape and skewness parameters respectively. Next, I used the assumed Gamma($\alpha = 10.4, \beta = 0.36$) FSW age distribution in South Africa from the Thembisa model (Johnson and Dorrington 2020) to calculate the implied ratio between the number of FSW and the sexually debuted population in each age group. I assumed these ratios in South Africa were applicable to every country, allowing calculation of the number of FSW by age group in all 13 countries. The results obtained are shown in Figure 4.2.

4.3.5 Results

Coverage assessment
Variance decomposition
Estimates

4.4 Prevalence and incidence by risk group

4.4.1 Disaggregation of Naomi estimates

I calculated HIV incidence λ_{iak} and number of new HIV infections I_{iak} stratified according to district, age group and risk group by linear disaggregation

$$I_{ia} = \sum_k I_{iak} = \sum_k \lambda_{iak} N_{iak} \tag{4.19}$$

$$= 0 + \lambda_{ia2} N_{ia2} + \lambda_{ia3} N_{ia3} + \lambda_{ia4} N_{ia4} \tag{4.20}$$

$$= \lambda_{ia2} \left(N_{ia2} + \mathrm{RR}_3 N_{ia3} + \mathrm{RR}_4(\lambda_{ia}) N_{ia4} \right). \tag{4.21}$$

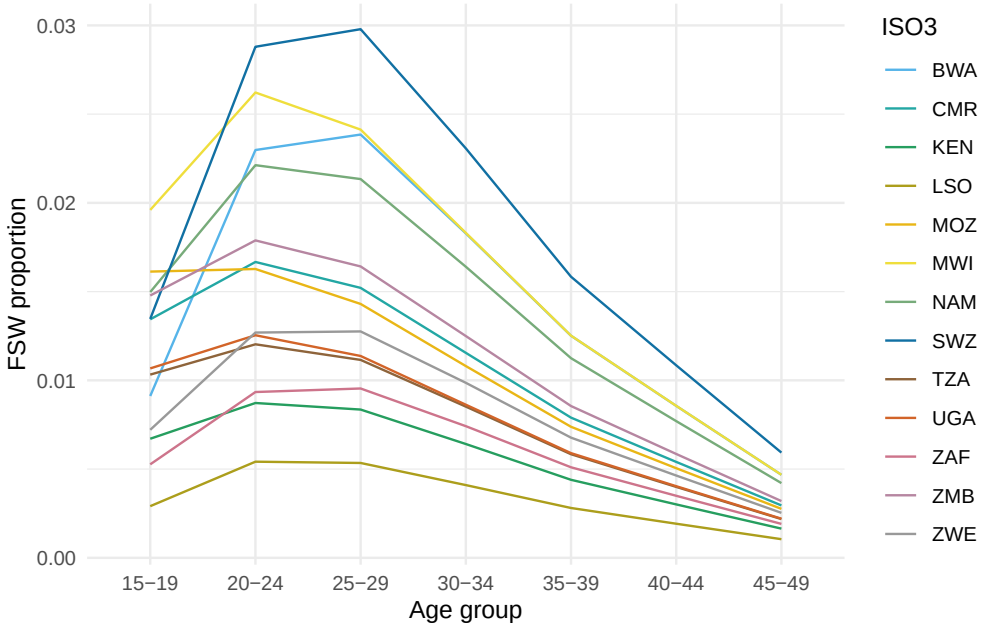

Figure 4.2: Proportion of FSW by age group (including the age groups 30-34, 35-39, 40-44 and 45-49) as produced by the disaggregation procedure.

Risk group specific HIV incidence estimates are then given by

$$\lambda_{ia1} = 0, \tag{4.22}$$

$$\lambda_{ia2} = I_{ia} / \left(N_{ia2} + \mathrm{RR}_3 N_{ia3} + \mathrm{RR}_4(\lambda_{ia}) N_{ia4} \right), \tag{4.23}$$

$$\lambda_{ia3} = \mathrm{RR}_3 \lambda_{ia2}, \tag{4.24}$$

$$\lambda_{ia4} = \mathrm{RR}_4(\lambda_{ia}) \lambda_{ia2}. \tag{4.25}$$

which I evaluated using Naomi model estimates of the number of new HIV infections $I_{ia} = \lambda_{ia} N_{ia}$, HIV infection risk ratios $\{\mathrm{RR}_3, \mathrm{RR}_4(\lambda_{ia})\}$, and risk group population sizes as above. The risk ratio $\mathrm{RR}_4(\lambda_{ia})$ was defined as a function of general population incidence. The number of new HIV infections are then $I_{iak} = \lambda_{iak} N_{iak}$.

4.4.2 Expected new infections reached

I calculated the number of new infections that would be reached prioritising according to each possible stratification of the population–that is for all $2^3 = 8$ possible combinations of stratification by location, age, and risk group. As an illustration, for stratification just by age, I aggregated the number of new HIV infections

and HIV incidence as such

$$I_a = \sum_{ik} I_{iak}, \tag{4.26}$$

$$\lambda_a = I_a / \sum_{ik} N_{iak}. \tag{4.27}$$

Under this stratification, individuals in each age group a are prioritised according to the highest HIV incidence λ_a. By cumulatively summing the expected infections, for each fraction of the total population reached I calculated the fraction of total expected new infections that would be reached.

4.4.3 Results

4.5 Discussion

I estimated the proportion of AGYW who fall into different risk groups at a district level in 13 sub-Saharan African countries. These estimates support consideration of differentiated prevention programming according to geographic locations and risk behaviour, as outlined in the Global AIDS Strategy. Systematic differences in risk by age groups, and variation within and between countries, explained the large majority of variation in risk group proportions. Changes over time were negligible in the overall variation in risk group proportions. The proportion of 15-19 year olds who are sexually active, and among women aged 20-29 years, norms around cohabitation especially varied across districts and countries. This variation underscores the need for these granular data to implement HIV prevention options aligned to local norms and risk behaviours.

I considered four risk groups based on sexual behaviour, the most proximal determinant of risk. Other factors, such as condom usage or type of sexual act, may account for additional heterogeneity in risk from sexual behaviour. However, I did not include these factors in view of measurement difficulties, concerns about consistency across contexts, and the operational benefits of describing risk parsimoniously. Sexual behaviour confers risk only when AGYW reside in geographic locations where there is unsuppressed viral load among their potential partners. I

did not include more distal determinants, such as school attendance, orphanhood, or gender empowerment, as I expect their effects on risk to largely be mediated by more proximal determinants. However, to effectively implement programming, it is crucial to understand these factors, as well as the broader structural barriers and limits to personal agency faced by AGYW. Importantly, programs must ensure that intervention prioritisation occurs without stigmatising or blaming AGYW.

Brugh et al. (**brugh2021characterizing**) previously geographically mapped AGYW HIV risk groups using biomarker and behavioural data from the most recent surveys in Eswatini, Haiti and Mozambique to define and subsequently map risk groups with a range of machine learning techniques. My work builds on Brugh et al. (**brugh2021characterizing**) by including more countries, integrating a greater number of surveys, and connecting risk group proportions with HIV epidemic indicators to help inform programming.

By considering a range of possible risk stratification strategies, I showed that successful implementation of a risk-stratified approach would allow substantially more of those at risk for infections to be identified before infection occurs. A considerable proportion of estimated new infections were among FSW, supporting the case for HIV programming efforts focused on key population groups (**baral2012burden**). There is substantial variation in the importance of prioritisation by age, location and behaviour within each country. This highlights the importance of understanding and tailoring HIV prevention efforts to country-specific contexts. By standardising our analysis across all 13 countries, I showed the additional efficiency benefits of resource allocation between countries.

I found a geographic delineation in the proportion of women cohabiting between southern and eastern Africa, calling attention to a divide attributable to many cultural, social, and economic factors. The delineation does not represent a boundary between predominately Christian and Muslim populations, which is further north. I also note that the high numbers of adolescent girls aged 15-19 cohabiting in Mozambique is markedly different from the other countries (**unicef**).

A model for risk group proportions

Our modelled estimates of risk group proportions improve upon direct survey results for three reasons. First, by taking a modular modelling approach, I integrated all relevant survey information from multiple years, allowing estimation of the FSW proportion for surveys without a specific transactional sex question. Second, whereas direct estimates exhibit large sampling variability at a district level, I alleviated this issue using spatio-temporal smoothing (Fig B in S2 Text). Third, I provided estimates in all district-years, including those not directly sampled by surveys, allowing estimates to be consistently fed into further analysis and planning pipelines (such as our analysis of risk group specific prevalence and incidence).

The final surveys included in our risk model model were conducted in 2018. I plan to update our analysis with more surveys as they become available, but do not anticipate that the risk group proportions will change substantially, as I found that they did not change significantly over time.

Our analysis focused on females aged 15-29 years, and could be extended to consider optimisation of prevention more broadly, accounting for the 0% of new infections among adults 15-49 which occur in women 30-49 and men 15-49. Estimating sexual risk behaviour in adults 15-49 would be a crucial step toward greater understanding of the dynamics of the HIV epidemic in sub-Saharan Africa, and would allow incidence models to include stratification of individuals by sexual risk.

Distribution of risk

- Connection to phylogenetic results from BDI
- About transmission rather than incidence
- Only age-sex structured not age-sex-behaviour
- Does not undermine my work

Community engagement

- CSO engagement
- Problem in Malawi with FSW

4.5.1 Limitations

Our analysis was subject to challenges shared by most approaches to monitoring sexual behaviour in the general population (**cleland2004monitoring**). In particular, under-reporting of higher risk sexual behaviours among AGYW could affect the validity of our risk group proportion estimates. Due to social stigma or disapproval, respondents may be reluctant to report non-marital partners (**nnko2004secretive**) or may bias their reporting of sexual debut (**zaba2004age**; **wringe2009comparative**; Nguyen and Eaton 2022). For guidance of resource allocation, differing rates of under-reporting by country, district, year or age group are particularly concerning to the applicability of our results; and, while it may be reasonable to assume a constant rate over space-time, the same cannot be said for age, where aspects of under-reporting have been shown to decline as respondents age (**glynn2011assessing**), suggesting that the elevated risks I found faced by younger women are likely a conservative estimate. If present, these reporting biases will also have distorted the estimates of infection risk ratios and prevalence ratios I used in our analysis, likely over-attributing risk to higher risk groups.

I have the least confidence in our estimates for the FSW risk group. As well as having the smallest sample sizes, our transactional sex estimates do not overcome the difficulties of sampling hard to reach groups. I inherent any limitations of the national FSW estimates (Stevens et al. 2022) which I adjust our estimates of transactional sex to match. Furthermore, I do not consider seasonal migration patterns, which may particularly affect FSW size. More generally, I did not consider covariates potentially predictive of risk group proportions (such as sociodemographic characteristics, education, local economic activity, cultural and religious norms and attitudes), which are typically difficult to measure spatially. Identifying measurable correlates of risk, or particular settings in which time-concentrated HIV risk occurs, is an important area for further research to improve risk prioritisation and precision HIV programme delivery.

The efficiency of each stratified prevention strategy depends on the ability of programmes to identify and effectively reach those in each strata. Our analysis of new infections potentially averted assumed a "best-case" scenario where AGYW of every strata can be reached perfectly, and should therefore be interpreted as illustrating the potentially obtainable benefits rather than benefits which would be obtained from any specific intervention strategy. In practice, stratified prevention strategies are likely to be substantially less efficient than this best-case scenario. Factors I did not consider include the greater administrative burden of more complex strategies, variation in difficulty or feasibility of reaching individuals in each strata, variation in the range or effectiveness of interventions by strata, and changes in strata membership that may occur during the course of a year. Identifying and reaching behavioural strata may be particularly challenging. Empirical evaluations of behavioural risk screening tools have found only moderate discriminatory ability (**jia2022risk**), and risk behaviour may change rapidly among young populations, increasing the challenge to effectively deliver appropriately timed prevention packages. This consideration may motivate selecting risk groups based on easily observable attributes, such as attendance of a particular service or facility, rather than sexual behaviour.

4.5.2 Conclusion

I estimated the proportion of AGYW aged 15-19, 20-24 and 25-29 years in four sexual risk groups at a district-level in 13 priority countries and analyzed the number of infections that could be reached by prioritisation based upon location, age and behaviour. Though subject to limitations, these estimates provide data that national HIV programmes can use to set targets and implement differentiated HIV prevention strategies as outlined in the Global AIDS Strategy. Successfully implementing this approach would result in more efficiently reaching a greater number of those at risk of infection.

Among AGYW, there was systematic variation in sexual behaviour by age and location, but not over time. Age group variation was primarily attributable to age

of sexual debut (ages 15-24). Spatial variation was particularly present between those who reported one cohabiting partner versus non-regular or multiple partners. Risk group proportions did not change substantially over time, indicating that norms relating to sexual behaviour are relatively static. These findings underscore the importance of providing effective HIV prevention options tailored to the needs of particular age groups, as well as local norms around sexual partnerships.

5

Fast approximate Bayesian inference

In this chapter I describe a novel Bayesian inference method I developed with the aim of facilitating fast and accurate inference for the Naomi small-area estimation model. This work builds on that of others, including Rue et al. (2009), Kristensen et al. (2016) and Stringer et al. (2021). I began working on this project in 2020, but did not make significant progress until I read Alex Stringer's work. We later began collaborating, including Alex supervising my visit to the University of Waterloo in 2022. The results are presented in Howes, Stringer, et al. (2023+).

Code for the analysis in this chapter is available from `athowes/naomi-aghq` and supported by the R package `inf.utils`.

5.1 Background

5.2 The Naomi model

5.3 Methods for inference

5.4 Malawi case-study

5.5 Discussion

6

Future work and conclusions

6.1 Strengths

6.1.1 Chapter 3

- I designed experiments to thoroughly compare models for spatial structure using tools for model assessment such as proper scoring rules and posterior predictive checks.

6.1.2 Chapter 4

- I estimated HIV risk group proportions for AGYW, enabling countries to prioritise their delivery of HIV prevention services.
- I analysed the number of new infections that might be reached under a variety of risk stratification strategies.
- I used `R-INLA` to specify multinomial spatio-temporal models via the Poisson-multinomial transformation. This includes complex two- and three-way Kronecker product interactions defined using the `group` and `replicate` options.

6.1.3 Chapter 5

- I developed a novel Bayesian inference method, motivated by a challenging and practically important problem in HIV inference.
- The method enables integrated nested Laplace approximations to be fit to and studied on a wider class of models than was previously possible.
- My implementation of the method was straightforward, building on the `TMB` and `aghq` packages, and described completely and accessibly in Howes, Stringer, et al. (2023+).

6.2 Future work

Avenues for future work include:

1. Extending the risk group model described in Chapter 4 to include all adults 15-49. This may involve modelling of age-stratified sexual partnerships (Wolock et al. 2021). Such a model would likely fall out of the scope of `R-INLA`, but would be possible to write with `TMB` and therefore amenable to the methods discussed in Chapter 5.
2. Speeding up the implementation of Laplace marginals using the matrix algebra approximations described in Wood (2020).
3. Evaluating the accuracy of deterministic Bayesian inference methods for a broader variety of extended latent Gaussian models.

6.3 Conclusions

- Modelling complex data, more often than not, pushes the boundaries of the statistical toolkit available.
- A challenge I encountered was the difficulty of implementing identical models across multiple frameworks with the aim of studying the inference method. Or, of a similarly fraught nature, comparing different models implemented in different frameworks with the aim of studying model differences. The

frequently asked questions section of the `R-INLA` website (Rue 2023) notes that "the devil is in the details". I have resolved this challenge by using a given `TMB` model template to fit models using multiple inference methodologies. The benefits of such a ecosystem of packages are noted by Stringer (2021). I particularly highlight the benefit of enabling analysts to easily vary their choice of inference method based on the stage of model development that they are in.

- To the best of my abilities, I have written this thesis, and the work described within it, in keeping with the principles of open science. I hope that doing so allows my work to be scrutinised, and optimistically built upon. This would not have been possible without a range of tools from the R ecosystem such as `rmarkdown` and `rticles`, as well as those developed within the MRC Centre for Global Infectious Disease Analysis such as `orderly` and `didehpc`.

Appendices

A

Spatial structure supplement

B

A model for risk group proportions supplement

C

Fast approximate Bayesian inference supplement

Works Cited

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". In: *Annals of the Institute of Statistical Mathematics* 43.1, pp. 1–20.

Cressie, Noel and Christopher K Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

DHS (2012). *Sampling and Household Listing Manual: Demographic and Health Surveys Methodology*.

Eaton, Jeffrey W et al. (2021). "Naomi: A New Modelling Tool for Estimating HIV Epidemic Indicators at the District Level in Sub-Saharan Africa". In.

Freni-Sterrantino, Anna, Massimo Ventrucci, and Håvard Rue (2018). "A note on intrinsic conditional autoregressive models for disconnected graphs". In: *Spatial and spatio-temporal epidemiology* 26, pp. 25–34.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. CRC press.

Hodgins, Caroline et al. (2022). "Population sizes, HIV prevalence, and HIV prevention among men who paid for sex in sub-Saharan Africa (2000–2020): A meta-analysis of 87 population-based surveys". In: *PLoS Medicine* 19.1, e1003861.

Howes, Adam (2022). *multi.utils: Utility functions for multi-agyw*. R package version 0.1.0.

Howes, Adam, Jeffrey W. Eaton, and Seth R. Flaxman (2023+). "Beyond borders: evaluating the suitability of spatial adjacency for small-area estimation". In.

Howes, Adam, Kathryn A. Risher, et al. (Apr. 2023). "Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa". In: *PLOS Global Public Health* 3.4, pp. 1–14. DOI: 10.1371/journal.pgph.0001731. URL: https://doi.org/10.1371/journal.pgph.0001731.

Howes, Adam, Alex Stringer, et al. (2023+). "Fast approximate Bayesian inference of HIV indicators using PCA adaptive Gauss-Hermite quadrature". In.

Johnson, L and RE Dorrington (2020). "Thembisa version 4.3: A model for evaluating the impact of HIV/AIDS in South Africa". In: *View Article*.

Kish, Leslie (1965). *Survey sampling*. 04; HN29, K5.

Kristensen, Kasper et al. (2016). "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software* 70.i05.

Lumley, Thomas (2004). "Analysis of Complex Survey Samples". In: *Journal of Statistical Software, Articles* 9.8, pp. 1–19. DOI: 10.18637/jss.v009.i08. URL: https://www.jstatsoft.org/v009/i08.

McCullagh, Peter and John A Nelder (1989). *Generalized linear models*. Routledge.

McElreath, Richard (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Nguyen, Van Kính and Jeffrey W. Eaton (2022). "Trends and country-level variation in age at first sex in sub-Saharan Africa among birth cohorts entering adulthood

between 1985 and 2020". In: *BMC Public Health* 22.1, p. 1120. DOI: 10.1186/s12889-022-13451-y. URL: https://doi.org/10.1186/s12889-022-13451-y.

Rue, Havard (2023). "'R-INLA' Project - FAQ". Accessed 23/01/2023. URL: https://www.r-inla.org/faq.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.

Saul, Janet et al. (2018). "The DREAMS core package of interventions: a comprehensive approach to preventing HIV among adolescent girls and young women". In: *PLoS One* 13.12, e0208167.

Simpson, Daniel et al. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors". In: *Statistical Science* 32.1, pp. 1–28.

Sørbye, Sigrunn Holbek and Håvard Rue (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling". In: *Spatial Statistics* 8, pp. 39–51.

— (2017). "Penalised complexity priors for stationary autoregressive processes". In: *Journal of Time Series Analysis* 38.6, pp. 923–935.

Stevens, Oliver et al. (2022). "Estimating key population size, HIV prevalence, and ART coverage for sub-Saharan Africa at the national level". In.

Stringer, Alex (2021). "Implementing Approximate Bayesian Inference Using Adaptive Quadrature". Statistics Graduate Student Research Day 2021, The Fields Institute for Research in Mathematical Sciences. URL: http://www.fields.utoronto.ca/talks/Implementing-Approximate-Bayesian-Inference-Using-Adaptive-Quadrature.

Stringer, Alex, Patrick Brown, and Jamie Stafford (2021). "Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models". In: *arXiv preprint arXiv:2103.07425*.

The Global Fund (2018). *The Global Fund Measurement Framework for Adolescent Girls and Young Women Programs*. Accessed 30/08/2021. URL: https://www.theglobalfund.org/media/8076/me_adolescentsgirlsandyoungwomenprograms_frameworkmeasurement_en.pdf.

UNAIDS (2021a). *2021 UNAIDS Global AIDS Update - Confronting Inequalities - Lessons for pandemic responses from 40 Years of AIDS*.

— (2021b). *Global AIDS Update 2021*. https://www.unaids.org/en/resources/documents/2021/global-aids-update. Accessed: June 2023.

Wolock, Timothy M et al. (June 2021). "Evaluating distributional regression strategies for modelling self-reported sexual age-mixing". In: *eLife* 10. Ed. by Eduardo Franco, Talía Malagón, and Adam Akullian, e68318. DOI: 10.7554/eLife.68318. URL: https://doi.org/10.7554/eLife.68318.

Wood, Simon N (2020). "Simplified integrated nested Laplace approximation". In: *Biometrika* 107.1, pp. 223–230.