Biometrics, 2024, 80(1), ujad015
https://doi.org/10.1093/biomtc/ujad015
Biometric Methodology

OXFORD

Inferring HIV transmission patterns from viral deep-sequence data via latent typed point processes

Fan Bu[1,2], Joseph Kagaayi[3], Mary Kate Grabowski[4], Oliver Ratmann[5], Jason Xu[6,*]

[1]Department of Biostatistics, University of California - Los Angeles, Los Angeles, CA 90024, United States, [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, United States, [3]School of Public Health, Makerere University, Kampala, Uganda, [4]School of Medicine, Johns Hopkins University, Baltimore, MD 21218, United States, [5]Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom, [6]Department of Statistical Science, Duke University, Durham, NC 27708, United States

[*]Corresponding author: Jason Xu, Department of Statistical Science, Duke University, Durham, NC 27708, United States (jason.q.xu@duke.edu).

ABSTRACT

Viral deep-sequencing data play a crucial role toward understanding disease transmission network flows, providing higher resolution compared to standard Sanger sequencing. To more fully utilize these rich data and account for the uncertainties in outcomes from phylogenetic analyses, we propose a spatial Poisson process model to uncover human immunodeficiency virus (HIV) transmission flow patterns at the population level. We represent pairings of individuals with viral sequence data as typed points, with coordinates representing covariates such as gender and age and point types representing the unobserved transmission statuses (linkage and direction). Points are associated with observed scores on the strength of evidence for each transmission status that are obtained through standard deep-sequence phylogenetic analysis. Our method is able to jointly infer the latent transmission statuses for all pairings and the transmission flow surface on the source-recipient covariate space. In contrast to existing methods, our framework does not require preclassification of the transmission statuses of data points, and instead learns them probabilistically through a fully Bayesian inference scheme. By directly modeling continuous spatial processes with smooth densities, our method enjoys significant computational advantages compared to previous methods that rely on discretization of the covariate space. We demonstrate that our framework can capture age structures in HIV transmission at high resolution, bringing valuable insights in a case study on viral deep-sequencing data from Southern Uganda.

KEYWORDS: Bayesian data augmentation; likelihood-based inference; marked spatial point processes; phylodynamics; Sub-Saharan Africa.

1 INTRODUCTION

As a decade-long global pandemic, the human immunodeficiency virus (HIV) has most severely affected Africa with 1 in every 25 adults living with the HIV virus, accounting for more than two-thirds of infections worldwide (Eisinger and Fauci, 2018; Fauci and Lane, 2020). International public health organizations target intervention efforts at populations most at risk of HIV acquisition and transmission (Glynn et al., 2001; Pettifor et al., 2008; Karim et al., 2010; Jewkes et al., 2010; Saul et al., 2018), motivating a better understanding of transmission patterns between different population groups (Wilson and Halperin, 2008).

To this end, this article introduces novel methods to infer transmission flows among different groups of individuals. We focus on modeling the age structure in heterosexual transmission patterns, representing transmission flows as latent surfaces in a plane with the source and recipient ages as the axes. Each transmission pair then becomes a point on this plane, with all transmission pairs corresponding to a realized point pattern. We expect that similar age groups exhibit similar behaviors, akin to modeling continuous spatial surfaces on a compact domain (Ji et al., 2009; Kutoyants, 2012). The key scientific challenge lies in the unobserved transmission pathways, where uncertainty exists regarding the occurrence and direction of transmissions between each pair. Answering the question "who infected whom?" is thus fundamental for learning population-level transmission flow.

To infer transmission pathways between individuals, we leverage outputs from modern phylogenetic analyses. Recent viral deep-sequencing pipelines have enabled estimation of transmission linkage and direction by inferring evolutionary relationships between individuals from multiple sampled viral sequences (Romero-Severson et al., 2016; Leitner and Romero-Severson, 2018; Wymant et al., 2018; Ratmann et al., 2019). These pipelines typically yield 2 summary scores indicating (1) the likelihood of shared transmission links among deep-sequenced individuals and (2) the probability of transmission in a specific direction (Wymant et al., 2018; Ratmann et al., 2019; Bbosa et al., 2020; Hall et al., 2021). However, these summary scores are imprecise and cannot definitively "prove" transmission between individuals (Zhang et al., 2021). Additionally, they typically provide only "maximum likelihood" phylogenetic structures without uncertainty quantification. Existing approaches for learning transmission flows usually apply heuristic thresholds to these summary scores, which disregard the varying strengths of phylogenetic evidence and also omit substantial fractions of data prior to analysis due to the thresholding. In

Received: February 23, 2023; Revised: October 9, 2023; Accepted: November 27, 2023

short, there is a substantial methodological gap for utilizing phylogenetic summaries in that (1) differential evidence confidence is neglected and (2) a large amount of data is discarded.

To address these limitations, this article proposes a coherent statistical model, which jointly learns from demographic information (such as gender and age) and phylogenetic evidence. We introduce a marked latent spatial process model on the age space, where each transmission pair of individuals (represented by their paired ages as coordinates on the space) is associated with "marks" that contain the phylogenetic summary scores. That is, each potential transmission pair is assigned a latent "type" that indicates the unknown transmission statuses (linkage and direction) as a random variable. The distribution of transmission flow between age groups and the distribution of the "marks" (phylogenetic scores) both depend on the latent type. We derive the likelihood of the complete model, so that basing inference in a data-augmented Bayesian framework then allows us to probabilistically learn the latent type for each potential transmission pair jointly with the parameters. In particular, the posterior formally quantifies the evidence strength for each pair of infected individuals—for example, a pair assigned an 85% posterior probability of linked transmission would contribute more to the learning of the flow surfaces compared to a pair with a 50% posterior linkage probability. Importantly, this joint modeling approach enables us to make use of substantially more data, as "low-confidence" pairs with lower phylogenetic scores reflecting weaker evidence of linkage or direction are downweighted in a data-driven manner rather than discarded.

In addition to making better use of the data and uncertainty, our latent spatial point process approach admits a more computationally efficient solution, owing to a continuous formulation of the transmission flow space. A common approach in past studies entails discrete grids based on prespecified age groups, such as 1 or 5-year age bands. These heuristic groupings can lead to computationally intensive analysis (Hyman et al., 1994; Heuveline, 2004; Sharrow et al., 2014). For instance, Xi et al. (2022) introduces a semiparametric Poisson model for flow counts on discrete age strata and other demographic attributes. The number of observed points is often considerably smaller than the number of cells in the discretized age space, leading to many structural zeros in the grid that demand considerable book-keeping and heavy computation for model smoothing downstream. Instead, our point pattern approach with a continuous underlying surface is at once more general and more computationally efficient, despite including latent variables. We will also show that the point process model borrows information in a two-way manner, leveraging the additional data to learn the transmission flows while using the learned flows to infer the point types.

Methodologically, we contribute a novel extension to existing statistical methodologies of spatial point processes. Spatial Poisson process (PP) models have been widely applied to the study of point-referenced 2D data (Banerjee et al., 2003; Huber, 2011; Cressie, 2015). These models have recently been extended to study point patterns that are latent or partially observed, where additionally observed "marks" associated with point patterns can be utilized to facilitate inferences (Vedel Jesen and Thorarinsdottir, 2007; Ji et al., 2009). To our knowledge, much of this existing work focuses on *one* set or type of spatial points, rather than a combination of multiple "types" of latent point patterns where the unobserved "type" is interpretable and of practical importance. Our framework bridges this methodological gap, leveraging additional information to infer latent "types" and extract meaningful structures under a marked-point process model. We provide a brief review of related prior works in Web Appendix A.2, and note that the statistical framework readily transfers to studying transmission dynamics for other infectious diseases (Paterson et al., 2015).

This paper is structured as follows: we provide an overview of the motivating data and develop the model framework in Section 2. A likelihood-based inference scheme is presented in Section 3. We then investigate inference accuracy under the proposed model and its ability to differentiate between competing transmission flow hypotheses on simulated data in Section 4. Next, we illustrate the efficacy of the proposed approach on demographic and HIV deep-sequence data from the Rakai Community Cohort Study (RCCS) in Southern Uganda in Section 5. Finally, we discuss the merits and future directions of our framework in Section 6.

2 DATA AND MODEL

2.1 Demographic and viral phylogenetic data of HIV infected individuals

Human immunodeficiency virus deep-sequence data were collected from blood samples of participants living with HIV in the RCCS, a longitudinal, population-based census and cohort study in Southern Uganda (Grabowski et al., 2017). Samples were obtained between August 2011 and January 2015 from 2652 individuals having an HIV viral load of >1000 copies/mL plasma and sufficient viral sequence read depth and length for deep sequence data analysis (Ratmann et al., 2019; Wymant et al., 2018). Detailed demographic, behavioral, and healthcare data were collected for all participants, including gender and age determined from self-reported birth dates and/or official personal documents (Grabowski et al., 2017).

Our model is motivated by a dataset consisting of 539 heterosexual pairs of HIV-infected RCCS participants, who are considered as phylogenetically possible transmission pairs. They were identified among all 3 515 226 pairwise combinations in all 2652 deep-sequence participants. We only considered pairs in 446 distinct subgraphs of viral deep-sequence phylogenies, interpreted as separate potential transmission networks with distinct viral introductions (Ratmann et al., 2019). We further investigated these transmission networks using the *phyloscanner* deep-sequence analysis pipeline (Wymant et al., 2018), which included transmission direction information, unlike typical phylogenetic cluster analyses (De Oliveira et al., 2017). This allowed us to eliminate heterosexual pairs that could not have occurred in acyclic transmission chains, resulting in 539 pairs for analysis. Figure 1 provides an illustration of this data set.

There are 2 main facets of the data. The first comprises the age of the 2 individuals in a pair at the midpoint of the observation period, denoted by $\mathbb{S} = \{\mathbf{s}_i = (a_{i1}, a_{i2})^T\}_{i=1}^N$ (sample size $N = 539$), where the 2D vector (a_{i1}, a_{i2}) records the male's age a_{i1}

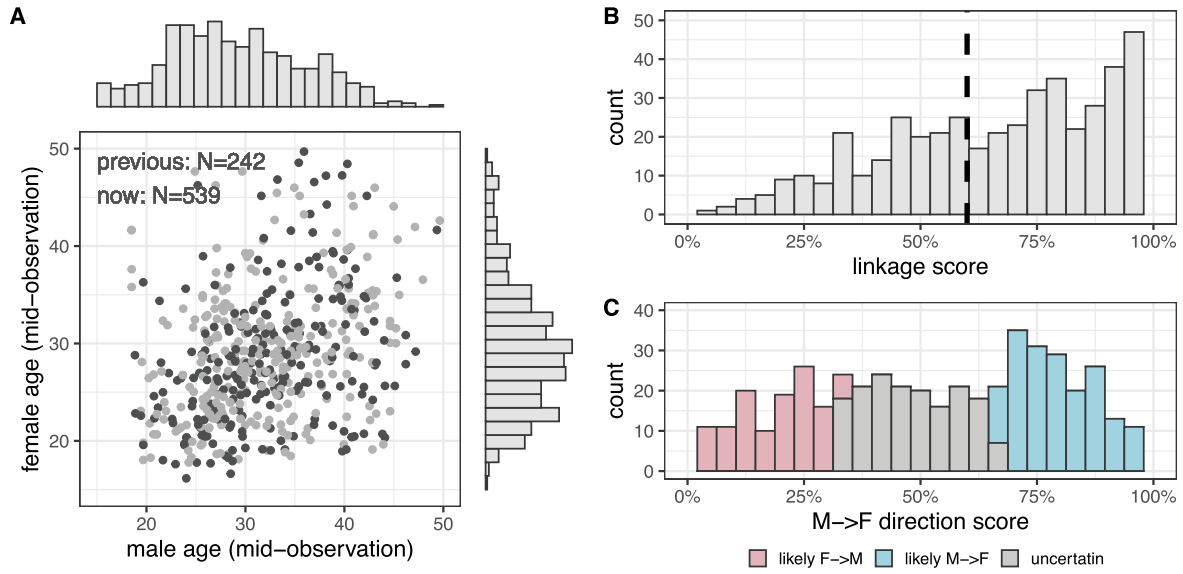

FIGURE 1 Point data and marks representing the output of HIV phylogenetic deep-sequence analysis on the sequence sample from Rakai, Uganda, from 2010 to 2015. (A) Paired ages of 539 heterosexual individuals who were inferred to be phylogenetically closely related with the HIV phylogenetic deep-sequence analysis using the *phyloscanner* software on HIV deep-sequence data from 2652 study participants of the Rakai Community Cohort Study in Southern Uganda, 2010–2015. The age of the individuals in the closely related pairs was calculated at the midpoint of the observation period, and the age of the men and of the women are shown on the x-axis and y-axis, respectively. Each data point is associated with two phylogenetic deep-sequence summary statistics in (0,1), the linkage score (ℓ_i), and the direction score (d_i) (see text). Points associated with high linkage and direction scores ($\ell_i \geq 0.6$ and $d_i \geq 0.67$) are shown in dark grey, and all other points are shown in light grey. Marginal histograms on the age of men and women are shown for all points. The typed point process model that we develop here aims to infer transmission flows using all data points rather than the highly likely "source-recipient" pairs shown in dark grey. (B) Histogram of the linkage scores across all data points. (C) Histogram of the direction scores across all data points. Direction scores $d_i \leq 1/3$ indicate high confidence in female-to-male transmission (shown in red), and direction scores $d_i \geq 2/3$ indicate high confidence in male-to-female transmission (shown in blue).

and female's age a_{i2} in the ith pair. Our model envisions these paired ages as observations from a spatial process describing the transmission structure. The second facet consists of 2 scores in the range of (0–1) that are outputs from phylogenetic analyses of HIV deep-sequencing data with *phyloscanner*. For each pair i, *phyloscanner* produces 2 scores—a linkage score and a direction score—by assessing the viral phylogenetic relationship of individuals in terms of the patristic distances and topological configurations of the viral reads in deep-sequence phylogenies, and then counting the observed patterns over sliding, overlapping genomic windows across the HIV genome (Wymant et al., 2018; Ratmann et al., 2019). The linkage score ℓ_i represents the posterior probability of the pair sharing a transmission link in the transmission process under a Binomial count model of window-specific linkage classifications (Ratmann et al., 2019). The direction score d_i, on the other hand, measures the posterior probability of transmission taking place from the male to the female in this pair under a similar count model. We collectively denote the phylogenetic scores for the ith pair by $\mathbf{x}_i = (\ell_i, d_i)^T$, and for brevity refer to the phylogenetic data as the "marks" associated with each of the points $\mathbf{s}_i$ for $i = 1, \ldots, N$.

The key data challenge is the unobserved transmission relationship between pairs of HIV-infected individuals. Even for a pair with phylogenetic evidence suggesting high probability to be linked through disease transmission (with a high ℓ_i score), we do not have direct knowledge about the transmission linkage and direction. Our model (described later) will, therefore, probabilistically characterize the likelihoods of pairwise transmission linkage and direction in a data-driven manner.

2.2 Substantial information loss in existing modeling approaches

Existing analyses attempt to address the unobserved pairwise transmission relationships by preclassifying the data points using heuristic thresholds on the phylogenetic summary scores (Xi et al., 2022; Ratmann et al., 2020; Hall et al., 2021). For example, a potential transmission pair i would be classified as a male-to-female transmission event if $\ell_i > 0.6$ and $d_i > 0.67$; similarly, another pair j would be taken as a female-to-male transmission if $\ell_j > 0.6$ and $d_j < 0.33$ (Xi et al., 2022). Such thresholding *a priori*, albeit procedurally simple, excludes a substantial proportion of data from the analysis, as "low-confidence" pairs are completely discarded. Graphically, Figure 1 shows that all data points with linkage scores falling to the left of the vertical line in panel B and all data with direction scores in the gray region of panel C would be discarded, resulting in only 242 out of 539 total data points (see panel A) retained for analysis. Further, such preclassification neglects differential strengths of phylogenetic evidence from data points classified as transmission events. Intuitively, we should have higher confidence about a likely transmission pair i with $\ell_i = 0.98$ to represent a transmission event, compared to another pair j with only $\ell_j = 0.61$.

However, this differential strength of evidence is not reflected in the existing methodology. Instead, all data points that are believed to be "high-confidence" pairs are treated as equal and exchangeable, which fails to fully exploit the rich information summarized by the phylogenetic scores. In the next section, we introduce our modeling framework that makes better use of the data by jointly leveraging phylogenetic evidence and demographic information.

2.3 The typed point process model

We now specify a general framework for inferring population-level transmission flows from the point pattern $\mathbb{S} = \{\mathbf{s}_i = (a_{i1}, a_{i2})^T\}_{i=1}^N$ with associated marks $\mathbf{x}_i = (\ell_i, d_i)^T$, that is, the phylogenetic transmission and direction scores. The marks reflect the strength of phylogenetic evidence regarding the unknown ground truth transmission relationship in each heterosexual pair. We introduce for each point a categorical latent random variable c_i that encodes 3 possible events: (i) transmission did not occur between the 2 individuals that define the point (denoted by $c_i = 0$), (ii) transmission occurred from the male to the female individual ($c_i = 1$), or (iii) transmission occurred from the female to the male individual ($c_i = -1$). For brevity, we refer to c_i as the latent "type" associated with each observed point $\mathbf{s}_i$ with mark $\mathbf{x}_i$. Intuitively, the modelling framework can be thought of as a typed spatial PP in that the intensity function of the process and the distribution of the marks both depend on the event type. This typed process then provides a generative model for marked point patterns. Though motivated by our particular application, the framework can be applied to many similar data sets of spatial point patterns of unknown types that are informed by marks associated with each point.

With this context in place, we begin by considering a spatial point pattern defined on a 2D space $\mathcal{S} \times \mathcal{S}$, where $\mathbb{S} = \{\mathbf{s}_i\}_{i=1}^N$ is the set of all points, and each point $\mathbf{s}_i = (s_{i1}, s_{i2})^T$ is represented as a point in this plane. We model the observed points in $\mathbb{S}$ as a realization of a 2D PP, $\mathbb{S} \sim PP(\boldsymbol{\lambda})$, on $\mathcal{S} \times \mathcal{S}$. Following Kottas and Sansó (2007) we decompose the intensity function $\boldsymbol{\lambda}$ into a scale component γ and a density function $f(\cdot)$, $\boldsymbol{\lambda}(\cdot) = \gamma f(\cdot)$, so that $f(\cdot)$ satisfies $\int_{S\times S} f(s_1, s_2) ds_1 ds_2 = 1$. This decomposition separates the intensity function into 2 terms, which are simpler to write out in the likelihood function and make inference computationally tractable.

We next model the density component $f(\cdot)$ as a mixture $f(\cdot) = \sum_{k\in\mathcal{K}} p_k f_k(\cdot)$, where p_k is the probability of points belonging to type k, and $f_k(\cdot)$ is the spatial density function for type k. In the context of our application, $\mathcal{S}$ is the continuous age of individuals under study, $\mathcal{S} = [15, 50)$. Each point $\mathbf{s}_i = (a_{i1}, a_{i2})^T$ corresponds to the ages of the 2 individuals forming a pair, ordered by gender and the latent types are $\mathcal{K} = \{-1, 0, 1\}$, corresponding to female-to-male transmission, no transmission and male-to-female transmission. For instance p_1 corresponds to the proportion of male-to-female transmission events among all pairs of individuals being considered, and $f_1(\cdot)$ corresponds to the 2D function that captures the across-age transmission pattern with male sources and female recipients.

2.4 Infinite Gaussian mixture models for the typed intensity functions

There are various choices to model the structure of the density functions $f_k(\cdot)$. To balance simplicity and flexibility, we choose a Dirichlet process (DP) Gaussian mixture model (DPGMM) consisting of infinitely many bivariate Gaussian components $f_k(\cdot)$. Specifically, for each point $\mathbf{s}_i$, if its type label $c_i = k$, then we have $\mathbf{s}_i | c_i = k \sim N(\theta_{ki}, \Sigma_{ki})$, $(\theta_{ki}, \Sigma_{ki}) \sim G_k$, and $G_k \sim DP(\alpha_k, G_0)$. Here G_k represents the (infinite) mixture of bivariate normal models for type k, and θ_{ki} and Σ_{ki} are the mean vector and covariance matrix for the bivariate Gaussian component that $\mathbf{s}_i$ belongs to.

In practice, Dirichlet process mixtures are often treated as a finite mixture but with a flexible number of components. Indeed, the above defined model may be expressed equivalently in terms of each density function $f_k(\cdot)$,

$$f_k(\cdot) = \sum_{h=1}^{H_k} w_{kh} \cdot \mathrm{dBVN}(\cdot; \theta_{kh}, \Sigma_{kh}), \tag{1}$$

where H_k denotes the number of "active" components, or total number of unique components generated by the DP, and each $(\theta_{kh}, \Sigma_{kh})$ is a unique Gaussian component for the type-k density. Here, $\mathrm{dBVN}(\cdot\,; \theta, \Sigma)$ denotes the probability density function of a bivariate normal distribution with mean θ and covariance Σ. Practically, we handle the H_k's by specifying a sufficiently large "maximum number of components", $H_{\max}$, and treat the DPGMM as a finite mixture model with $H_{\max}$ components (in our analysis $H_{\max} = 10$ seems to be sufficient); this technique has been discussed in (Ji et al., 2009). We present a sensitivity analysis on the choice of $H_{\max}$ in Web Appendix C.3.

2.5 Model for type-dependent marks

We next describe how the observed point patterns and their associated marks are connected through the typed point process model. Assuming that marks $\mathbf{x}_i$ associated with each point $\mathbf{s}_i$ can provide information on the true event type c_i of each point, we model the distribution of marks $\mathbf{x}_i$ conditional on type c_i. In our application, the marks $\mathbf{x}_i = (\ell_i, d_i)^T$ are 2D vectors with entries taking values in (0,1), and we assume the following type-dependent distribution for $\mathbf{x}_i$ conditional on the type value $c_i = k$ ($k \in \{-1, 0, 1\}$):

$$\begin{aligned} p(\mathbf{x}_i \mid c_i = k) = \phi_k((\ell_i, d_i)^T) = dN(\mathrm{logit}(\ell_i); \tilde{\mu}_{\ell,i}(k), \sigma_\ell^2) \\ \times dN(\mathrm{logit}(d_i); \tilde{\mu}_{d,i}(k), \sigma_d^2). \end{aligned} \tag{2}$$

Here, $dN(\cdot\,; \mu, \sigma^2)$ denotes a univariate normal density function with mean μ and variance σ^2, and $\mathrm{logit}(x)$ denotes the logit transformation on $x \in (0, 1)$. We further specify type-dependent normal means $\tilde{\mu}_{\ell,i}(k)$ and $\tilde{\mu}_{d,i}(k)$ as

$$\begin{aligned} \tilde{\mu}_{\ell,i}(k) &= \mu_\ell \mathbb{1}[k \neq 0]; \\ \tilde{\mu}_{d,i}(k) &= \mu_d \mathbb{1}[k = 1] + \mu_{-d} \mathbb{1}[k = -1]. \end{aligned} \tag{3}$$

Intuitively, a larger linkage score ℓ_i indicates stronger evidence for a transmission link, and a larger direction d_i indicates higher confidence for a male-to-female transmission. Thus, with $\mu_\ell > 0$, the first part in (3) implies that ℓ_i likely exceeds 0.5 for a real transmission event ($c_i \neq 0$); similarly, with $\mu_{-d} < 0 < \mu_d$, the

second part in (3) implies that d_i is probably larger than 0.5 for a male-to-female event ($c_i = 1$) but smaller than 0.5 for a female-to-male event ($c_i = -1$). Note this design uses the property $\text{logit}(0.5) = 0$.

2.6 The complete data likelihood

From the descriptions above, the 2 key components in the model—the spatial process and the marks distribution—are linked through the latent types c_i. Conditional on the latent type $c_i = k$, a point i contributes the term $\gamma p_k f_k(\mathbf{s}_i) \times \phi_k(\mathbf{x}_i)$ to the data likelihood, and therefore the likelihood simply require multiplying all such terms for $i = 1, 2, \ldots, N$. With *complete* data, which include the coordinates of all N points in the set $\mathbb{S}$, the observed signals $\mathbf{x}_i$ *as well as the type* c_i for all $i = 1, \ldots, N$, we can write down the complete data likelihood function $L(\Theta; \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\})$. Model parameters include $\Theta = \{\gamma, \mathbf{p}, \boldsymbol{\mu}, \sigma_\ell^2, \sigma_d^2, \{(\theta_{kh}, \Sigma_{kh})\}, \{\alpha_k\}\}$, where $p = (p_{-1}, p_0, p_1)^T$, and $\boldsymbol{\mu} = (\mu_\ell, \mu_d, \mu_{-d})^T$).

$$
\begin{aligned}
&L(\Theta; \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\}) \\
&= \gamma^N \frac{e^{-\gamma}}{N!} \prod_{k \in \mathcal{K}} \prod_{i:c_i=k} p_k f_k(\mathbf{s}_i) \phi_k(\mathbf{x}_i) \qquad (4) \\
&= \prod_{i=1:N} dN(\text{logit}(\ell_i); \tilde{\mu}_{\ell,i}(c_i), \sigma_\ell^2) dN \\
&\quad \times (\text{logit}(d_i); \tilde{\mu}_{d,i}(c_i), \sigma_d^2) \\
&\quad \times \gamma^N \frac{e^{-\gamma}}{N!} \prod_{k \in \mathcal{K}} \prod_{i:c_i=k} \left(p_k \sum_{h=1}^{H_k} w_{kh} \text{BVN} \right. \\
&\quad \left. \times ((s_{i1}, s_{i2}); \theta_h, \Sigma_h) \right). \qquad (5)
\end{aligned}
$$

3 BAYESIAN INFERENCE WITH DATA AUGMENTATION

We employ an efficient data-augmented Bayesian inference scheme (Tanner and Wong, 1987; Van Dyk and Meng, 2001) to learn the unknown parameters Θ in the proposed typed point process model from observed data. Inference would be straightforward if all aspects of the model were observable, giving access to the complete data likelihood specified in (5). That is, if the c_i's were known, the terms corresponding to the spatial point process and marks would completely factorize in the likelihood function, as shown in (5). Parameter inference would reduce to standard procedures for Dirichlet Process Gaussian mixture models (Rasmussen, 1999).

However, the type c_i for each data point i is not observed in our data setting. Inference through the marginal likelihood based only on observed data would entail an intractable high-dimensional integration step. Instead, we propose a data augmentation scheme that expands the target posterior to include the unobserved c_i's as unknowns, which allows us to exploit the convenient expression of the complete data likelihood in (5). That is, our algorithm samples from the joint posterior density of the model parameters Θ together with the unobserved types c_i's: $p(\Theta, \{c_i\}|\{\mathbf{x}_i\}, \{\mathbf{s}_i\}) \propto L(\Theta; \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\}) p_0(\Theta)$. Here $p_0(\Theta)$ denotes the joint prior distribution for parameters Θ. The data-augmented inference framework is employed through a Bayesian Markov chain Monte Carlo (MCMC) sampler. The algorithm can be roughly divided into 2 main components in each iteration: (1) sample or update parameters Θ conditioned on configurations of the $\{c_i\}$'s from the conditional posterior distribution $p(\Theta|\{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\})$ and (2) sample c_i for each i given values of Θ from $p(c_i|\Theta, \mathbf{x}_i, \mathbf{s}_i)$, utilizing the complete data likelihood in (5). To improve the efficiency of the MCMC sampler, we prescribe conjugate or semiconjugate priors whenever possible, enabling straightforward Gibbs sampling by exploiting the full conditional posterior densities available for almost all parameters. We provide details on all prior choices and sampling steps in Web Appendix A.2.

4 SIMULATION STUDIES

In this section, we validate our proposed framework through synthetic experiments. As is relevant to our application, we assess the accuracy of the typed point process model in recovering simulated patterns of HIV transmission flows between men and women across different ages. Additionally, we explore the model's capability to infer these flows with increasingly smaller sample sizes. We benchmark its performance with a subset model that preclassifies potential transmission pairs using an existing approach described in Section 2.2, where spatial components are still learned with the DPGMM framework. We will focus on 2 contemporary questions of of particular epidemiological interest, and investigate the model's ability to recover and differentiate between 2 competing scenarios.

First, recent HIV transmission flow studies have shown that more infections tend to originate from men than from women (Bbosa et al., 2020; Ratmann et al., 2020; Hall et al., 2021). This motivates us to investigate the proposed framework's ability to accurately recover parameters and distinguish between two scenarios: one where men drive 50% of infections ("MF 50-50") and another where men drive 60% of infections ("MF 60-40"). We conduct experiments with five different sample sizes: $N =$ 100, 200, 400, 600, and 800. For each scenario and sample size, we generate 100 independent datasets and fit the typed point process model to each.

Second, there is significant interest in the age distribution of male sources of HIV infections, especially to adolescent and young women aged 15-24 given their high incidence rates (Risher et al., 2021). We explore whether the model can differentiate between 2 scenarios: one where younger men (around 25 years old) contribute 60% of infections in women aged 15–24, while older men (around 35 years old) contribute 30% and other men contribute 10% ("SAME AGE") and another scenario where younger men contribute 30%, older men contribute 60%, and other men contribute 10% ("DISCORDANT AGE"). We vary sample sizes, generate 100 simulated datasets for each scenario, and fit the model as previously described (see Web Appendix B for complete details).

Figure 2 summarizes our findings from the two sets of simulation studies, where we use darker colors to represent

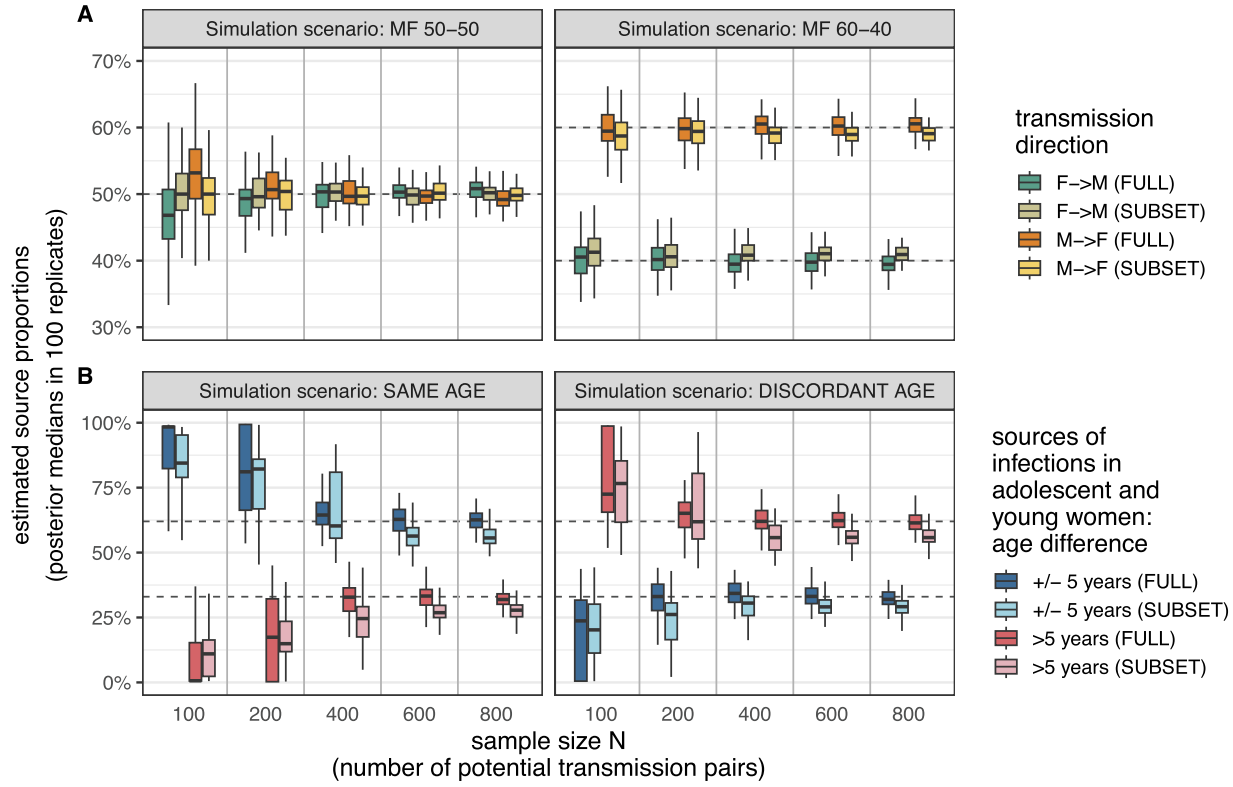

FIGURE 2 Performance of the typed point process model in recovering simulated transmission flow patterns. Key parameters are estimated using the typed point process model ("FULL", darker colors) and a subset model that uses an existing approach to pre-classify point types ("SUBSET", lighter colors) on each of the 100 simulated data sets under each scenario. (A) Boxplot of the posterior mean estimate of transmissions from men in 100 replicate simulations for the MF 50-50 (left panel) and MF 60-40 scenarios (right panel). Throughout, the dashed lines mark the true values that underpin the simulated data. The *x*-axis shows the sample size of simulated data points, which represent the number of phylogenetically closely related pairs of individuals identified through phylogenetic deep-sequence analyses. (B) Boxplot of the posterior mean estimate of transmissions from men of similar age (shown in red) and older age (shown in blue) to infection in adolescent and young women aged 15–24 in 100 replicate simulations for the "SAME AGE" and "DISCORDANT AGE" scenarios. As before, the dashed lines mark the true values that underpin the simulated data and the x-axis shows results for different sample sizes.

results from our proposed model ("FULL") and lighter colors for the subset model with preclassification ("SUBSET"). The top panel in Figure 2 shows results from the male-female simulation experiments (MF 50-50 and MF 60-40). We find that our proposed model is able to successfully distinguish between the 2 competing epidemiological scenarios and produce estimates for gender-related proportions that are accurate within a $\pm 5\%$ error margin for sample sizes $N \geq 200$. The bottom panel in Figure 2 illustrates our findings on the age-specific sources experiments (SAME AGE and DISCORDANT AGE). This is a substantially more difficult inference problem because the target quantities relate to a smaller subgroup of the entire source population. With a small sample (N in the range 100–200), the relative relationship between the 2 proportions is inferred correctly in all experiments. However, the actual quantitative estimates and differences can deviate from the truth, in this case overestimated. For sample sizes of $N \geq 400$, the quantitative estimates become satisfactorily accurate. The observed overestimation is likely due to the parsimony induced by the Dirichlet Process priors that are known to prefer assigning data points to the largest existing clusters when there are not enough data to admit a new mixture model component. As a result, more points tend to be attributed to the component with the highest weight when N is small; this effect is mitigated as N increases.

Between the 2 sets of simulation experiments, the proposed model consistently produces more accurate estimates for both gender-related and age-related proportions compared to the existing pre-classifying approach. We note that in the "MF 50-50" scenario the 2 models are comparable, but this is because data are simulated with symmetric distribution for the direction scores, and in this case the subset model benefits from preclassification heuristics that happen to match the ground truth. Additional analyses on simulated experiments that include numerical convergence and mixing analyses as well as Bayesian coverage analyses, are provided in Web Appendix B.

5 CASE STUDY

We next apply our point process model to demographic and population-based HIV deep-sequence data from the RCCS (Ratmann et al., 2019, 2020). Our goal is to reconstruct transmission flows by gender and continuous age between 15

TABLE 1 Proportions and numbers of inferred event types.

Type	Full analysis with latent event types		Subset analysis with fixed event types	
	p_k	N_k	p_k	N_k
Male-to-female transmission (k = 1)	46.3% (39.4%, 53.1%)	244 (207, 279)	35.7%	188
Female-to-male transmission ($k = -1$)	35.0% (28.5%, 42.0%)	184 (150, 221)	29.5%	155
No transmission ($k = 0$)	18.6% (9.4%, 29.2%)	98 (49, 154)	34.8%	183

Posterior mean estimates with 95% credible intervals.

and 50 years. This case study is challenging as standard phylodynamic analyses using HIV consensus sequences struggle to infer flow patterns for more than a few age groups, typically limited to 5-year or 10-year age bands (Scire et al., 2020; Bbosa et al., 2020; Xi et al., 2022). We analyze all pairs of potential phylogenetically related individuals without heuristic preclassification, resulting in a dataset of 526 pairs representing potential transmissions upon excluding few pairs with very weak evidence. For comparison, we also implement an analysis with additional heuristic filtering as in (Xi et al., 2022), retaining only 367 "high-confidence" transmission pairs. We refer to the former approach as "full analysis," where we use latent variables to account for uncertainties in lower confidence pairs, and the latter as "subset analysis," using fixed types from the preprocessing. More details are provided in Web Appendix C.

Computational advantages. We are able to infer both the male-to-female and female-to-male transmission flow surfaces, f_1 and f_{-1}, and the unknown event types $\{c_i; i = 1, \ldots, N\}$ without issues in numerical convergence or mixing. We have performed diagnostics to ensure MCMC convergence, along with sensitivity analysis on the choice of priors and hyper-parameters, with details discussed in Web Appendix C.3. On a laptop with an 8-core CPU, a 30-minute runtime allows us to obtain 4000 MCMC samples after burn-in and thinning without parallelization. This is a considerable improvement in computational efficiency compared to using the semiparametric Poisson count model on 1-year age discretized bands (Xi et al., 2022), where 4000 total iterations require about 30 hours.

5.1 Learning latent event types results in more data used for inference of transmission flows

Compared to the subset analysis, the full analysis attributes more pairs of individuals to type 1 (male-to-female transmission) and to type −1 (female-to-male transmission). We compare the posterior probability p_k for each type k under each analysis in Table 1, as well as the number of pairs N_k attributed to each type upon multiplying p_k by the sample size. Notably, the full analysis learns a significantly lower p_0 (proportion of "no transmission"). This suggests that by inferring event types instead of heuristically preclassifying them, our proposed approach is utilizing more phylogenetic information from potential pairs.

5.2 Inferred age of male and female sources of HIV transmission

The typed point process model allows us to estimate the sources of male and female HIV infections continuously in terms of age, as shown in Figure 3A. This approach provides more flexibility compared to prior work, as we can summarize results at any desired resolution instead of being restricted to 5-year or 10-year age bands (De Oliveira et al., 2017; Le Vu et al., 2019; Bbosa et al., 2020; Scire et al., 2020). Coarse age discretization can obscure differences in important transmission modes from a public health perspective Xi et al. (2022). Figure 3A clearly answers the question, "which age groups contribute most to HIV transmission at the population level?" The colored solid curves represent the inferred age distributions of male and female sources in the full analysis, while the dark dashed lines display those under the subset analysis.

Male sources tend to be consistently older than female sources. The estimated 50% highest posterior density intervals (HDIs) under the full analysis are ages [25.4, 34.3] for male sources and [20.4, 29.0] for female sources. In comparison, the analogous intervals under the subset analysis are [26.9, 35.4] and [21.4, 29.3], respectively. Notably, the full analysis does not decrease estimation uncertainty in key epidemiological quantities. Instead, it better reflects the actual uncertainty by explicitly accounting for uncertainty in the event types underlying each data point.

Both analyses reveal a characteristic peak in transmission from men around age 30, with a long, pronounced tail of transmissions from men older than 40. In contrast, the age distribution of female sources differs qualitatively, as the probability of transmissions from women declines rapidly with age. These observations are more strongly supported by including latent variables in the full analysis—the entire age distribution of the female sources is slightly shifted toward younger ages compared to the inferred distribution under fixed types.

To gain further insight into age-specific transmission dynamics, we consider the age profile of the transmitting partners. In Figure 4, the recipient ages are color-coded, and the wave on the y-axis shows the posterior median contribution of transmissions to recipients of that age. Figure 4A illustrates that the age profile of sources has a characteristic shape for each recipient age group—they are not simply shifted versions of one age profile. Figure 4B demonstrates how the superposition of the age-structured transmission dynamics results in the overall source profile that marginalizes out the age of the recipients.

5.3 Transmissions to and from adolescent and young women

Next, we focus on adolescent and young women between age 15 and 24. Specifically, we wish to understand the age distribution of their male sources, and in turn, that of male recipients for whom these women are the sources of infection. Insights can provide further evidence to HIV programs that aim to reduce infections in this critical age group (Glynn et al., 2001; Karim et al.,

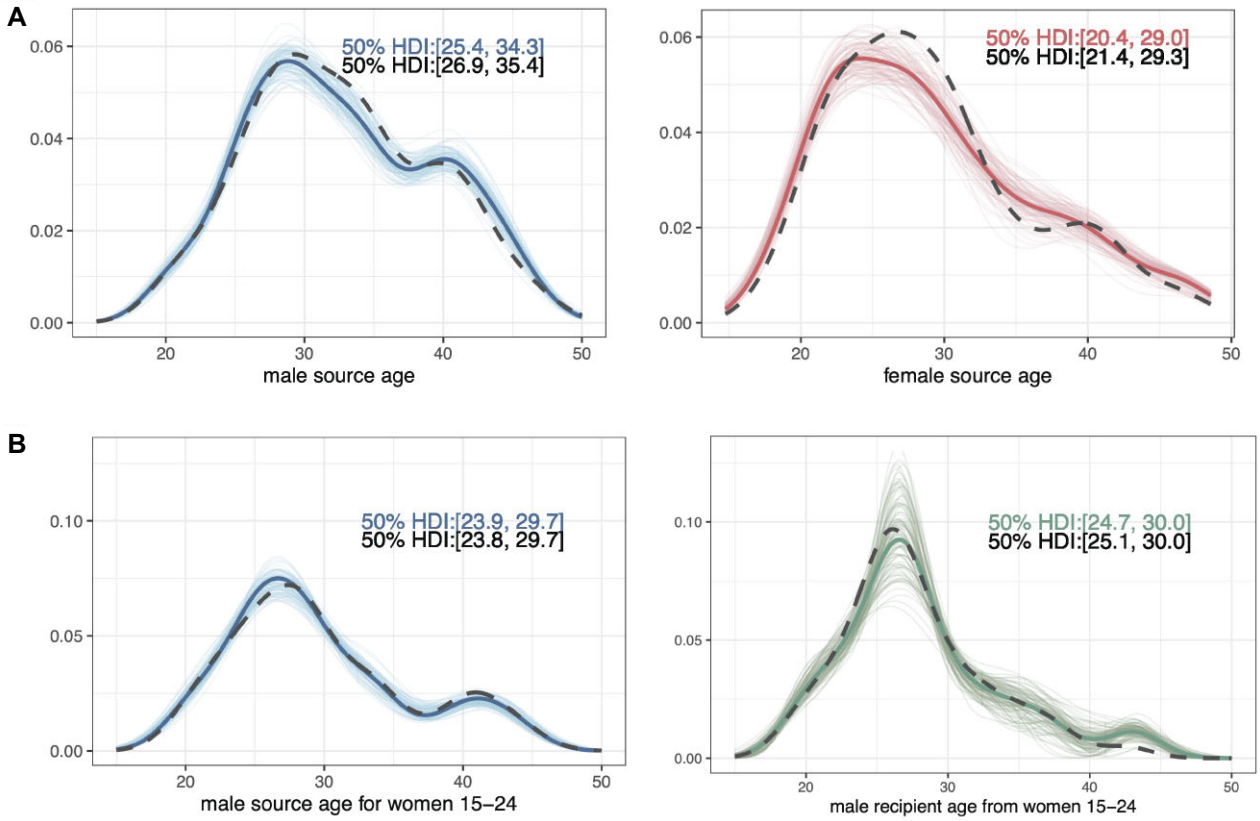

FIGURE 3 (A) Age distributions of male and female sources of HIV infections in Rakai, Uganda during the 2011-2015 observation period. The left panel shows the estimated age of the male sources and the right panel shows the estimated age of the female sources. (B) Age distributions of male sources and recipients of HIV infections in women aged 15-24. The left panel shows the age distribution of male sources and the right panel characterizes the age distribution of male recipients, for women aged 15-24. In each panel, the colored lines represent density curves of the age of sources/recipients for 100 posterior samples from the inferred, smooth transmission flow intensity surface of the typed point process model in the full analysis with latent event types. The thicker curve indicates the posterior mean density curve. The black dashed curve illustrates the posterior mean density curve in the subset analysis with fixed event types. A total of 50% highest density intervals (HDIs) are marked in text, with colored text indicating the HDIs inferred in the full analysis and black text indicating the HDIs inferred in the subset analysis.

2010; Jewkes et al., 2010), such as the DREAMS program (Saul et al., 2018) within the US President's Emergency Plan for AIDS Relief (PEPFAR) (Oliver, 2012).

Figure 3B presents our findings on inferred age distributions of male sources (left panel) and male recipients of transmissions from young women (right panel). Colored curves represent results from the full analysis with latent event types, while black curves correspond to the subset analysis. The majority of male sources fall within the 24-30 year range, with a notable subgroup of older men over 35. Conversely, male recipients of transmission from young women lie more strongly in the 24-30 year range, with a smaller subgroup of recipients over 35. This suggests a primary transmission pathway starting with transmission from men approximately 5-8 years older than the young women, who then transmit the virus to men 5-8 years older than themselves. Another pathway involves transmission from men 10-15 years older than young women, who then transmit the virus to men 5-8 years older than themselves, rather than 10-15 years older. These findings emphasize the significance of HIV prevention programs in Sub-Saharan Africa that target adolescent and young women, who are vulnerable to HIV from a broad age range of male sources (UNAIDS, 2018).

5.4 Comparing full analysis with latent event types to subset analysis using fixed types

We close by examining the finer details of the inferred age- and gender-specific transmission flows between analyses. In the subset analysis with fixed event types (Figure 5A), only a fraction of all data points are utilized to learn the latent age structure associated with each type. All data points carry the same weight. In the full analysis using latent types (Figure 5B), all potential pairs contribute to the learning of the latent age structure associated with each event type, and the contribution of each data point is weighted by its associated posterior type probability (illustrated with the color shades of each point).

Overall, the surfaces of age-specific male-to-female and female-to-male transmission flows appear similar under both approaches, as shown in Figure 5. This similarity is unsurprising given the age distributions depicted in Figure 3A. However, there are notable differences in the inferred latent surface corresponding to no transmission between the full and subset analyses. Furthermore, in the full analysis, data points involving women around age 20 and men around age 30 are more strongly attributed to male-to-female transmission, while data points

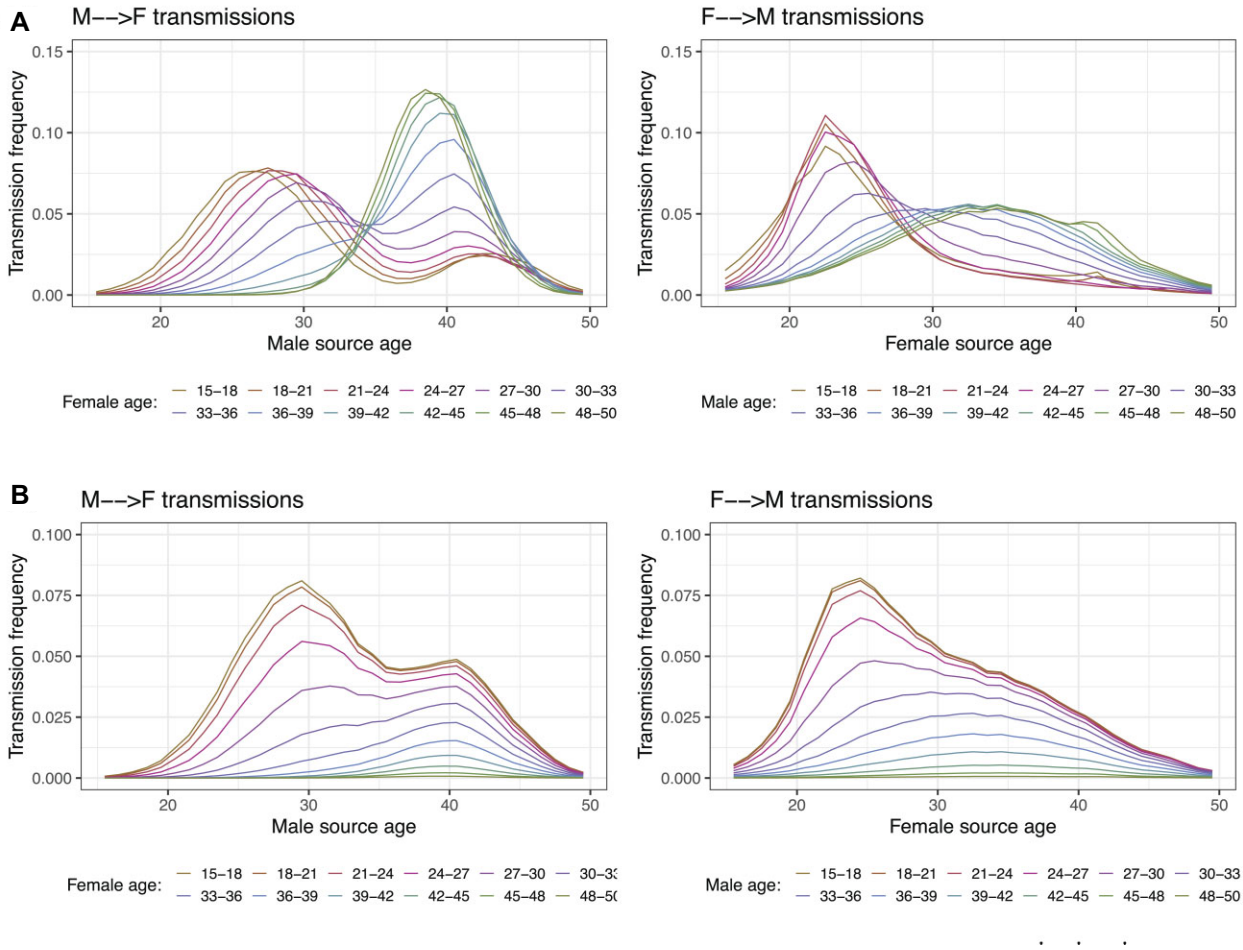

FIGURE 4 (A) Age distributions of sources for recipients in different 3-year age groups. Each curve represents the learned relative frequencies of sources responsible for transmissions within each recipient age group. (B) Marginal age distributions of sources shown as "'stacked" curves of age source distributions for each recipient age group (in 3-year age bands). In subplot 4(b), the "marginal" age distributions are shown by stacking up the frequencies of sources for each recipient age group. Left column shows age distributions for male sources in male-to-female transmissions for different female recipient age groups. Right column shows age distributions for female sources in female-to-male transmissions for different male recipient age groups.

involving women around age 45 are more strongly attributed to female-to-male transmission. The latter observation implies a potential extension to our approach where we could incorporate findings from prior studies as informed priors: women in this age are found to have lower HIV viral loads at the population level than men, making them less infectious on average (Grabowski et al., 2017; Rodger et al., 2019).

6 DISCUSSION

We propose a hierarchical typed point process model to learn disease transmission flows from phylogenetically reconstructed transmission pair data. Our novel approach includes a computationally efficient Bayesian inference algorithm that probabilistically learns the unobserved event types despite the large number of latent parameters and partially informative data. As illustrated in simulations and the case study, our framework efficiently utilizes more data and quantifies evidence strength in a data-driven manner, incorporating the uncertainties of phylogenetic summary outputs.

Using a continuous spatial process allows us to address epidemiologically important questions with fine-grained age-specific transmission patterns, overcoming the limitations of existing coarse age band analyses due to technical or computational constraints. Unlike discrete spatial treatments used in aggregated count data analyses, our typed point process approach avoids heavy computational burdens associated with Gaussian Markov random fields or similar models requiring intense matrix operations (Rue and Held, 2005). Given point-level data, our continuous spatial model simplifies computation significantly.

Our framework exhibits certain limitations that suggest future research directions. First, we consider only age and gender in modeling transmission flows, but we could consider including other individual covariates that may impact transmissibility as covariates in the Poisson processes or marks distributions (Hu and Bradley, 2018). Second, the normal mixture model assumes spherical proximity, which could be relaxed by extending to more flexible, non-normal mixture components such as the bivariate Beta kernel (Kottas and Sansó, 2007). Last but not

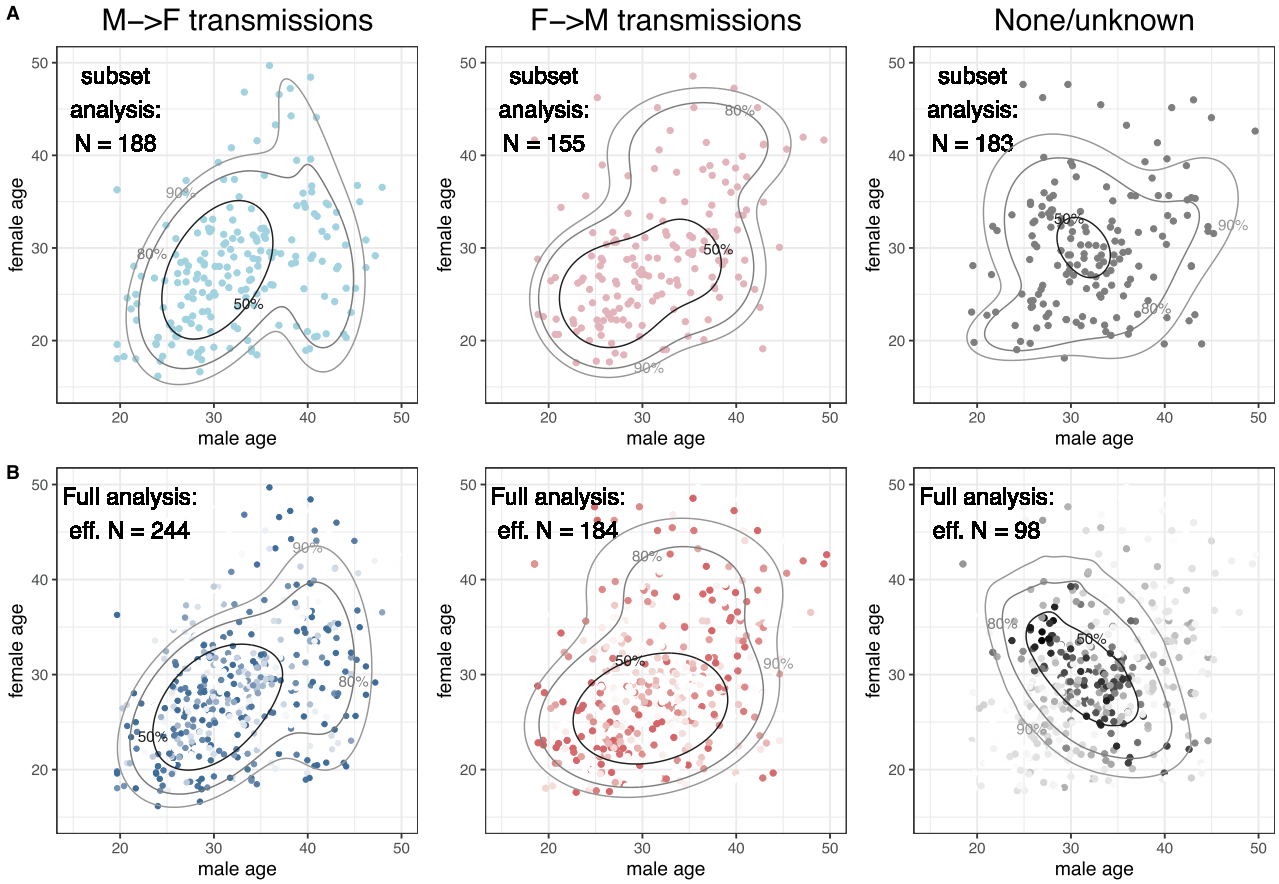

FIGURE 5 Comparison of the inferred age structure in transmission flows in the full analysis with latent event types versus the subset analysis with fixed event types. (A) Results in the subset analysis with fixed event types. Source-recipient pairs that were preclassified by event type (dots) are shown along the posterior median estimate of 50%, 80%, and 90% highest probability regions of transmission flows (contours). The number of data points attributed to each type is indicated in the top left corner. (B) Results in the full analysis with latent event types. Source-recipient pairs (dots) are shown by posterior event type probabilities (color intensity) along the posterior median estimate of 50%, 80%, and 90% highest probability regions of transmission flows (contours). The "effective" number ("eff. N") of data points attributed to each type (posterior mean estimate of N_k as in Table 1) is indicated in the top left corner.

the least, we have assumed conditional independence between transmission pairs, which could be generalized in order to leverage relevant dependency structures by introducing clusters or subgroups among infected individuals.

ACKNOWLEDGMENTS

We thank the staff, investigators, and participants of the Rakai Community Cohort Study and PANGEA-HIV who made this work possible. We thank Mike West for helpful discussions and Xiaoyue Xi for helpful comments and data preprocessing.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

All Web appendices, supplementary figures, and tables referenced in Sections 1-5, along with anonymized data and code in `R` and `Python` to replicate results in Sections 4 and 5, are available with this paper at the Biometrics website on Oxford Academic.

FUNDING

This work was partially supported by the National Science Foundation (DMS-2030355, DMS-2230074, and PIPP-2200047), the National Institutes of Health (R01 AI153044 and R01 AI155080-01A1), the Engineering and Physical Sciences Research Council (EP/X038440/1), and the Bill and Melinda Gates Foundation (OPP1175094, OPP1084362).

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

Anonymized data and code underlying this article are available at https://github.com/fanbu1995/HIV-transmission-PoissonProcess under the GNU General Public License version 3.0. The deep-sequence phylogenies and basic individual-level data analysed during the current study are available in the Dryad repository (DOI: 10.5061/dryad.7h46hg2). HIV-1 reads are

available upon reasonable request through the PANGEA consortium; see https://www.pangea-hiv.org. Additional individual-level data are available on reasonable request to the Rakai Health Sciences Program: see https://www.rhsp.org.

REFERENCES

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Bbosa, N., Ssemwanga, D., Ssekagiri, A., Xi, X., Mayanja, Y., Bahemuka, U. et al. (2020). Phylogenetic and demographic characterization of directed HIV-1 transmission using deep sequences from high-risk and general population cohorts/groups in Uganda. *Viruses*, 12, 331.

Cressie, N. (2015). *Statistics for Spatial Data*, John Wiley & Sons.

De Oliveira, T., Kharsany, A. B., Gräf, T., Cawood, C., Khanyile, D., Grobler, A. et al. (2017). Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *The Lancet HIV*, 4, e41–e50.

Eisinger, R. W. and Fauci, A. S. (2018). Ending the HIV/AIDS pandemic. *Emerging Infectious Diseases*, 24, 413.

Fauci, A. S. and Lane, H. C. (2020). Four decades of HIV/AIDS—much accomplished, much to do. *New England Journal of Medicine*, 383, 1–4.

Glynn, J. R., Caraël, M., Auvert, B., Kahindo, M., Chege, J., Musonda, R. et al. (2001). Why do young women have a much higher prevalence of HIV than young men? a study in Kisumu, Kenya and Ndola, Zambia. *Aids*, 15, S51–S60.

Grabowski, M. K., Serwadda, D. M., Gray, R. H., Nakigozi, G., Kigozi, G., Kagaayi, J. et al. (2017). HIV prevention efforts and incidence of HIV in Uganda. *New England Journal of Medicine*, 377, 2154–2166.

Hall, M., Golubchik, T., Bonsall, D., Abeler-Dorner, L., Limbada, M., Kosloff, B. et al. (2021). Demographic characteristics of sources of HIV-1 transmission in Zambia. *medRxiv*, https://doi.org/10.1101/2021.10.04.21263560, preprint: not peer reviewed.

Heuveline, P. (2004). Impact of the HIV epidemic on population and household structure: the dynamics and evidence to date. *AIDS (London, England)*, 18, S45.

Hu, G. and Bradley, J. (2018). A Bayesian spatial–temporal model with latent multivariate log-gamma random effects with application to earthquake magnitudes. *Stat*, 7, e179.

Huber, M. (2011). Spatial point processes. *Handbook of Markov Chain Monte Carlo*, 253–278.

Hyman, J. M., Li, J. and Stanley, E. A. (1994). Threshold conditions for the spread of the HIV infection in age-structured populations of homosexual men. *Journal of Theoretical Biology*, 166, 9–31.

Jewkes, R. K., Dunkle, K., Nduna, M. and Shai, N. (2010). Intimate partner violence, relationship power inequity, and incidence of HIV infection in young women in South Africa: a cohort study. *The Lancet*, 376, 41–48.

Ji, C., Merl, D., Kepler, T. B. and West, M. (2009). Spatial mixture modelling for unobserved point processes: Examples in immunofluorescence histology. *Bayesian Analysis*, 4, 297.

Karim, Q. A., Sibeko, S. and Baxter, C. (2010). Preventing HIV infection in women: a global health imperative. *Clinical Infectious Diseases*, 50, S122–S129.

Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137, 3151–3163.

Kutoyants, Y. A. (2012). *Statistical Inference for Spatial Poisson Processes*, Springer Science & Business Media.

Le Vu, S., Ratmann, O., Delpech, V., Brown, A. E., Gill, O. N., Tostevin, A. et al. (2019). HIV-1 transmission patterns in men who have sex with men: insights from genetic source attribution analysis. *AIDS Research and Human Retroviruses*, 35, 805–813.

Leitner, T. and Romero-Severson, E. (2018). Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nature Microbiology*, 3, 983–988.

Oliver, M. (2012). The US president's emergency plan for AIDS relief: Gendering the intersections of neo-conservatism and neo-liberalism. *International Feminist Journal of Politics*, 14, 226–246.

Paterson, G. K., Harrison, E. M., Murray, G. G., Welch, J. J., Warland, J. H., Holden, M. TG. et al. (2015). Capturing the cloud of diversity reveals complexity and heterogeneity of mrsa carriage, infection and transmission. *Nature Communications*, 6, 1–10.

Pettifor, A. E., Levandowski, B. A., MacPhail, C., Padian, N. S., Cohen, M. S. and Rees, H. V (2008). Keep them in school: the importance of education as a protective factor against HIV infection among young South African women. *International Journal of Epidemiology*, 37, 1266–1273.

Rasmussen, C. (1999). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12.

Ratmann, O., Grabowski, M. K., Hall, M., Golubchik, T., Wymant, C., Abeler-Dörner, L. et al. (2019). Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nature Communications*, 10, 1–13.

Ratmann, O., Kagaayi, J., Hall, M., Golubchick, T., Kigozi, G., Xi, X. et al. (2020). Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *The Lancet HIV*, 7, e173–e183.

Risher, K., Cori, A., Reniers, G., Marston, M., Calvert, C., Crampin, A. et al. (2021). Age patterns of HIV incidence in eastern and southern Africa: a collaborative analysis of observational general population cohort studies. *LANCET HIV*, 8.

Rodger, A. J., Cambiano, V., Bruun, T., Vernazza, P.,Collins S., Degen, O. et al. (2019). Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy (PARTNER): final results of a multicentre, prospective, observational study. *The Lancet*, 393, 2428–2438.

Romero-Severson, E. O., Bulla, I. and Leitner, T. (2016). Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 113, 2690–2695.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman and Hall/CRC.

Saul, J., Bachman, G., Allen, S., Toiv, N., Cooney, C. and Beamon, T. (2018). Determined resilient empowered AIDS-free mentored and safe (DREAMS): What is the core package and why now. *PLOS One*, 13, e0208167.

Scire, J., Barido-Sottani, J., Kühnert, D., Vaughan, T.G. and Stadler, T. (2022). Robust Phylodynamic Analysis of Genetic Sequencing Data from Structured Populations. *Viruses*, 14, 1648. https://doi.org/10.3390/v14081648.

Sharrow, D. J., Clark, S. J. and Raftery, A. E. (2014). Modeling age-specific mortality for countries with generalized HIV epidemics. *PloS one*, 9, e96447.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–540.

UNAIDS (2018). Miles to go: closing gaps, breaking barriers, righting justice. Global AIDS update 2018. UNAIDS, https://www.unaids.org/sites/default/files/media_asset/miles-to-go_en.pdf [Accessed 16 June 2019].

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10, 1–50.

Vedel Jesen, E. B. and Thorarinsdottir, T. L. (2007). A spatio-temporal model for functional magnetic resonance imaging data–with a view to resting state networks. *Scandinavian Journal of Statistics*, 34, 587–614.

Wilson, D. and Halperin, D. T. (2008). "Know your epidemic, know your response": a useful approach, if we get it right. *The Lancet*, 372, 423–426.

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M. et al. (2018). PHYLOSCANNER: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular Biology and Evolution*, 35, 719–733.

Xi, X., Spencer, S. E., Hall, M., M. K., Grabowski, Kagaayi, J. and Ratmann, O. (2022). Inferring the sources of HIV infection in Africa from deep sequence data with semi-parametric Bayesian Poisson flow models. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 71, 517–40.

Zhang, Y., Wymant, C., Laeyendecker, O., Grabowski, M. K., Hall, M., Hudelson, S. et al. (2021). Evaluation of phylogenetic methods for inferring the direction of human immunodeficiency virus transmission: HIV Prevention Trials Network (HPTN) 052. *Clinical Infectious Diseases*, 72, 30–37.

Received: February 23, 2023; Revised: October 9, 2023; Accepted: November 27, 2023