JMIR Publications
The Leading eHealth Publisher

Estimation of HIV Prevalence at the ZIP Code-Level in Atlanta, Georgia: Bayesian Prediction Modeling Using Passive Surveillance Data and Social Determinants of Disease Spreading.

Enrique Saldarriaga, Anirban Basu

Submitted to: JMIR Public Health and Surveillance
on: June 26, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Estimation of HIV Prevalence at the ZIP Code-Level in Atlanta, Georgia: Bayesian Prediction Modeling Using Passive Surveillance Data and Social Determinants of Disease Spreading.

Enrique Saldarriaga[1] PhD, MSc, BSc; Anirban Basu[1] PhD

[1]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute University of Washington Seattle US

Corresponding Author:
Enrique Saldarriaga PhD, MSc, BSc
The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute
University of Washington
1959 NE Pacific St.
HSB, H-375. Box357630
Seattle
US

Abstract

Background: Better information at the ZIP Code-level has the potential to enhance interventions targeting, identify treatment gaps, and optimize resources utilization. Currently there are no methods designed to estimate undiagnosed HIV cases at jurisdictions smaller than counties.

Objective: This study aims to predict the number of undiagnosed HIV cases at the ZIP Code-level in Atlanta, Georgia, based on publicly available information.

Methods: The CDC reports both passive surveillance (PS) and estimated total (MS) HIV cases for selected counties as part of the Ending of the HIV Epidemic initiative. We employed a Bayesian hierarchical model to: 1) Model MS as random draws from a Poisson distribution with mean equal to the true total HIV cases in the county. 2) A Binomial model for PS arising from the true denominator, with mean P, known as the ascertainment probability. 3) Use a logistic fractional model to allow P to be dependent on socio-economic determinants of HIV extracted from the American Community Survey. These determinants were chosen through a feature selection algorithm. The prediction model was tested out-of-sample on Georgia counties. Finally, we combined zip-code-level covariate data with the posterior predictive distribution of the logit coefficients to predict the mean P at zip-code-level. Final estimates were spatially-smoothed and aggregated to county-level for secondary validations.

Results: The county-level model showed good mixing properties and predictive accuracy. The mean ascertainment probability calibrated to the ZIP Code-level varied from 78.4% (95% credibility interval: 24.4%-99.3%) to 93.8% (95%CI: 80.6%-99.8%). Further, the predicted undiagnosed HIV cases ranged between 12 (95%CI: 6-19; ZIP Code 30322) to 1,603 (95%CI 1,209-1,968; ZIP Code 30318).

Conclusions: Our findings provide a more detailed understanding of the risk profile of the city, in particular regarding the heterogeneity and concentration of cases within the city, and therefore a more complete picture of the transmission risk. This information could be leveraged to better identify underserved communities, better targeting the delivery of prevention and treatment services, and overall increase the efficiency in the control of the HIV epidemic. Furthermore, our methodological approach can be applied to other cities in the country, to obtain a more detailed depictions of its HIV risk-profile and complement passive surveillance efforts. Clinical Trial: NA

(JMIR Preprints 26/06/2023:50310)
DOI: https://doi.org/10.2196/preprints.50310

Preprint Settings

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Estimation of HIV Prevalence at the ZIP Code-Level in Atlanta, Georgia: Bayesian Prediction Modeling Using Passive Surveillance Data and Social Determinants of Disease Spreading.

Enrique M. Saldarriaga, PhD, [1§] Anirban Basu, PhD[1]

[1]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, University of Washington, 1959 NE Pacific St, H-375, Seattle, WA 98195

Precis: Estimation of HIV prevalence at the ZIP Code-level

[§]**Corresponding author:**
Enrique M Saldarriaga
Email: emsb@uw.edu;
Telephone: +1 206.573.3133

E-mail addresses:
EMS: emsb@uw.edu
AB: basua@uw.edu

Word count:
Abstract: 429/450
Main text: 3,812

Abstract

Introduction. Local data has the potential to aid in the identification of treatment gaps, enhance interventions targeting, and consequently increase the efficiency of resources utilization for HIV prevention and control. While passive surveillance offers data of diagnosed cases at several geographical levels, it is subject to selection and participation bias, and the methods developed to estimate undiagnosed cases are only available up to the county level. ZIP Code data is a limitedly explored tool to better understand the HIV risk profile of a city and focus public health efforts accordingly. This study aims to predict the number of undiagnosed HIV cases at the ZIP Code-level in Atlanta, Georgia, based on publicly available information.

Methods. The CDC reports both passive surveillance and estimated total HIV cases for selected counties as part of the Ending of the HIV Epidemic (EHE) initiative. We employed a Bayesian hierarchical model to 1) Model total cases as random draws from a Negative Binomial distribution with mean equal to the true total HIV cases in the county incorporating its measurement error in the distribution's hyperparameters. 2) A Binomial model for passive surveillance cases arising from the true denominator, with mean P, known as the ascertainment probability. 3) Use a logistic fractional model to allow P to be dependent on socio-economic determinants of HIV extracted from the American Community Survey. These determinants were chosen through a feature selection algorithm. The prediction model was assessed out-of-sample on Georgia counties. Finally, we combined zip-code-level covariate data with the posterior predictive distribution of the logit coefficients to predict the mean P at zip-code-level. Final estimates were spatially-smoothed and aggregated to county-level for secondary validations.

Results. The county-level model showed good mixing properties and predictive accuracy. The mean ascertainment probability calibrated to the ZIP Code-level varied from 78.4% (95% credible interval: 24.4%-99.3%) to 93.8% (95%CI: 80.6%-99.8%). Further, the predicted undiagnosed HIV cases ranged between 12 (95%CI: 6-19; ZIP Code 30322) to 1,603 (95%CI 1,209-1,968; ZIP Code 30318).

Conclusions. Our findings provide a more detailed understanding of the risk profile of the city, regarding the heterogeneity and concentration of cases within the city, and therefore a more complete picture of the relative burden of HIV across ZIP Codes. While this information is relevant at the city level, the most actionable information could be obtained at the county level, where Local Health Departments could use our findings to identify underserved areas and allocate resources accordingly. Furthermore, our methodological approach can be applied to complement the information obtained from passive surveillance, especially when more resource-intensive approaches are not available or are unfeasible to employ.

Keywords: Passive Surveillance, Prediction modeling, Feature selection algorithm, Bayesian model, Spatial Smoothing, HIV Risk-Profile

Introduction

Since the introduction and expansion of the AIDS surveillance case definition in the United States in 1993 and the introduction of highly active antiretroviral therapy in 1995, the incidence and mortality of HIV have dropped by 69% and 48%, respectively.[1] However, this progress has not been evenly distributed among demographic groups. In 2018, 51% of all new diagnoses in the United States occurred in the South, even though only 38% of the population resided in this area.[2] The rate of people with HIV (PWH) in the South (372 per 100,000 people) is twice that of the mid-West (180 per 100,000) and 18% higher than the national average (313 per 100,000).[3] The HIV epidemic has grown disproportionately in the South compared to other regions in the country. A phenomenon driven by large urban centers [4] and concentrated on ethnic and racial minorities [3,5,6]. This is a consequence of long-standing structural inequalities across non-independent demographic and socioeconomic factors.[7,8]

To achieve the goals established by the Ending of the HIV Epidemic (EHE) initiative [9], which aims to reduce new HIV infections by 90% by 2030, it is imperative to design solutions to maximize access to care, particularly among historically marginalized demographic groups.[10] Local estimates are a tool that procures means to identify highly infected areas, treatment gaps, to improve interventions' targeting, and consequently enhance resource allocation efficiency. Interventions based on geographically aggregated information could overlook differences in needs within a given area, leading to inefficient resource allocation. This issue could be worsened in contexts where health outcomes are heterogeneous within a jurisdiction, for which there is evidence in the prevalence of HIV.[11] ZIP Codes are frequently used as the unit of analysis in public health research due to the granularity of information they provide and its established linkage with social determinants of health.[12,13] This is a limitedly explored opportunity to understand better the HIV risk profile of a city or county at a more nuanced geographic level to focus public health efforts accordingly. Better information at the ZIP Code-level has the potential to enhance interventions targeting to increase access to care, identify treatment gaps, and optimize resources utilization while improving health outcomes for PWH.[10,14]

A critical barrier in using ZIP Code-level data to inform decision-making is the absence of estimates of undiagnosed HIV cases at such level. While the HIV passive surveillance system allows local health departments to gather diagnosed cases at the ZIP Code level,[15,16] this data is susceptible to participation bias and under-ascertainment because not all diseased cases will be diagnosed.[17,18] On average, the proportion of undiagnosed HIV cases with passive surveillance at the county-level is 16.4% (95% confidence interval 15.7, 17.2).[19] However, the Centers for Disease Control and Prevention (CDC) surveillance reports show that this value can vary depending upon sex, age groups, race and ethnicity, and region of residence.[20] Access to healthcare services driven by multi-factorial inequalities are at the core of differences in HIV diagnosis.[8,21]

To obtain estimates of all prevalent HIV cases and level of under-ascertainment, the CDC uses the CD4 depletion model which estimates the time since infection as a function of the CD4-count in diagnosed cases and extrapolates using patterns of testing and reporting.[22] The model is applied to the national and state-level data in the US and the county-level data for the 50 counties in the Phase I EHE initiative.[20] However, these estimations had not been made at geographic areas below the county-level, likely because the methodology used requires data on testing and diagnosis patterns representative at the ZIP Code-level.[19]

We employed a novel method to predict total, diagnosed and undiagnosed, prevalent HIV cases at the ZIP Code-level based on passive surveillance data, social determinants of HIV spreading, and prevalence estimates at the county level. We focused on Atlanta Metropolitan Area given its high heterogeneity in HIV diagnoses distribution [11] and healthcare access disparities [23].

Methods

Data

We extracted county-level information from the CDC's HIV Surveillance Report Vol 25 [20], which contains the passive surveillance data (diagnosed cases) and estimated mean and 95% confidence interval of prevalent total (diagnosed and undiagnosed) HIV cases among people aged 13 years and above living with a known HIV infection in 2018. To ensure consistency, we excluded the information from the District of Columbia and Puerto Rico to include only counties. The sample had 48 observations, including four counties from Georgia: Cobb, DeKalb, Fulton, and Gwinnett counties. The training set included all but the four counties from Georgia, which were used as the test set. At the ZIP Code-level, we collected the count of prevalent diagnosed HIV cases, identified via passive surveillance, among persons aged 13 years and above in 2018, for the Atlanta Metropolitan Area, available through the Emory University's visualization tool, AIDSVu.[16] AIDSVu obtains this information via an agreement with the health departments overseeing HIV surveillance in the city to ensure the data follows standard procedures for privacy protection and geographical accuracy.[16] The analytical data set included 132 out of the 133 in Atlanta, with one ZIP Code (30334) being excluded because it reported less cases than the cut-off for data suppression to ensure non-identifiability of cases.

At both the County and ZIP Code-levels, we use the US Census Bureau's American Community Survey (ACS) 5-years to extract all potential predictors of the probability of ascertainment using a prospective selection based on literature review.[4,24,25] Selected variables included: age, gender, race and ethnicity, income, schooling, urbanicity, employment status, insurance coverage, wealth inequality, and vehicle ownership. We examined these variables to determine the need for transformations and interactions. Using the ACS ensured consistency in the collection methods at both the county- and ZIP Code-levels.

We explored four approaches to initial covariate selection: 1. All potential predictors. 2. We conducted a RIDGE regression on all potential predictors and excluded variables with evidence of multicollinearity (i.e., a variance inflation factor of 7 or more) or low explanatory power (i.e., variables whose coefficient had an absolute value of 1e-6 or less). The cut-off points were chosen somewhat arbitrarily with the intent of minimizing the possibility of excluding potentially relevant variables. 3. Covariates based on prior knowledge, including variables with the highest expectation of being associated with the probability of ascertainment. 4. The union of non-zero coefficients at optimum penalization from sets 2 and 3. For each initial approach, we used a LASSO regression with a penalization parameter tuned via leave-one-out cross-validation and estimate the Mean Squared Error (MSE) associated to each set. We chose the model with the lowest MSE to implement the prediction model.

Prediction Model

The probability of ascertainment of the passive surveillance system in a jurisdiction (i.e., county or ZIP Code) is the ratio of diagnosed cases over total cases. For a ZIP Code, denoted by i, this probability is expressed as $P_i = pscases_i / mcases_i$, where P is the proportion of detected cases, $pscases$ is the diagnosed cases obtained via passive surveillance, and $mcases$ is the total count of PWH, diagnosed and undiagnosed cases. This identity can be rewritten as $mcases_i = pscases_i / P_i$. The denominator in that equation is unknown at the ZIP Code-level, but it can be predicted.

We define a hierarchical Bayesian model at the county level, c. In the model, $pscases$ arises from a Binomial distribution with a true denominator $mcases$ and probability, P. The county-specific diagnosed cases, $mcasesMU$, was the target data to estimate $mcases$. We modeled $mcasesMU$ as arising from a Poisson distribution whose rate parameter, λ, distributed Gamma using an alternative parameterization of the Negative Binomial distribution. The parameters of λ, were calculated using the estimated true count of cases, $\widehat{mcases}$, and the relative standard error, RSE, associated to $mcasesMU$, because this variable was obtained through statistical modeling and therefore carried measurement error.[20] Furthermore, we allowed P to be dependent on county-level socioeconomic determinants of HIV using a logit link function. The model is defined as follows:

$$pscase\,s_c \quad Binomial\left(probability=P_c, trials=\widehat{mcases}_c\right)$$

$$mcasesM\,U_c \quad Poisson\left(rate=\lambda_c\right) \;\; \lambda_c \quad Gamma\left(shape=a_c, scale=b_c\right) \;\; a_c=\widehat{mcases}_c/b_c$$

$$b_c=\widehat{mcases}_c\cdot RS\,E_c^2 \;\; \widehat{mcase\,s}_c \quad Gamma(0.1,1e5)$$

$$logit\left(P_c\right)=\alpha_c+\beta\cdot X_c$$

$$\beta \quad Cauchy(0,2.5)$$

$$\alpha_c \quad Normal\left(mean=amu, SD=\tau\right)$$

$$amu \quad Normal(0,100)$$

$$\tau \quad Uniform\left(min=0, max=100\right)$$

Variables $pscase\,s_c$, $mcasesM\,U_c$, $RS\,E_c^2$, and X_c were observed data at the county level. X is a matrix of covariates with prediction power over P. $mcasesMU$, followed a Gamma-Poisson mixture distribution that allowed the model to acknowledge the overdispersion in the data (i.e., the mean and standard deviation are not equal as assumed by the Poisson distribution) and to use both the mean of total cases, $\widehat{mcases}$, and standard error, RSE, individually for each county through the hyperparameter λ_c. P was modeled with a hierarchical logistic model with a random intercept defined by a central tendency ,amu, and deviations ,tau, in addition to shared coefficients for the covariates, β, that captured the effect of the social determinants of HIV spreading.

Most hyperparameters (i.e., α, amu, τ, and β) used diffuse priors to allow the data to influence the estimation as much as possible.[26] On the other hand, β follows a Cauchy (i.e., Student-T distribution with 1 degree of freedom) distribution, aligned with best practices for priors in logistic regressions.[27] For $\widehat{mcases}$ we used a more informative prior to improve the definition of the binomial model, given that both of its parameters are estimated in the model.

The model was fitted to the training set. We used the trace plots and Gelman-Rubin's convergence statistic (R-hat) to assess the mixing properties of the model, overlap across chains, and low autocorrelation across iterations, with a cutoff of 1.1 to determine good mixing.[28] The analysis was conducted in the R software using *RJAGS*.[29]

Prediction and Validation

We used the posterior predictive distribution of the county-level ascertainment probability, P, coefficients in combination with the ZIP Code-level data of the covariates to predict the probability of ascertainment and the distribution of $\widehat{mcases}$ at the ZIP Code-level.

Further, to account for the spatial correlation structure of the data we fitted the estimated $\widehat{mcases}$ from all ZIP Codes to a spatial random-effect intrinsic conditional autoregressive (ICAR) model for a continuous outcome using the R-INLA package.[30,31] This model considers the spatial structure of the data and smooths the outcome geographically by incorporating information from neighboring ZIP Codes. The model is defined as follows:

$$\widehat{mcase\,s_i} = \alpha_0 + S_i + e_i + \epsilon_i \quad \widehat{mcase\,s_i} \sim Poisson\left(E_i \exp\left(\alpha_0 + S_i + e_i\right)\right) \quad e_i \vee \sigma_e^2 \sim_{iid} Normal\left(0, \sigma_e\right)$$

$$S_i \vee \sigma_S \sim_{iid} ICAR\left(\sigma_S^2\right)$$

Where i denotes ZIP Code, and $\widehat{mcase\,s_i}$ is the predicted values of total cases for each ZIP Code; E is the offset in the Poisson distribution, i.e., the population in each ZIP Code; α_0 is the average relative risk shared across areas; e_i is the non-spatial (unstructured) random effect for each ZIP Code and σ_e^2, its variance; S_i is the spatial random effect that follows ICAR parameterized with its variance, σ_S^2; ϵ is the measurement error. This analysis yielded mean and 95% credible intervals (95%CI) prevalence for each ZIP Code.

We validated the prediction in two ways. First, county-to-county: We estimated the mean and 95%CI of total cases for the four Georgian counties using the posterior predictive distribution of the logit coefficients in combination with their respective data and compared our predictions to the CDC's reported total cases. Second, ZIP Code-to-county: We aggregated the spatially smoothed estimates at the ZIP Code-level to estimate the county-level HIV cases, and compared our predictions to the county-level estimates. We used the estimates of the Department of Housing and Urban Development to determine the proportion of a ZIP Code (combining residents and territory) that falls into each county and adjust the number of cases accordingly.[32] We quantified prediction accuracy using the Mean Absolute Error (MAE).

Ethics

This study obtained an ethics review waiver from the University of Washington IRB. This is a secondary analysis of publicly available information where no primary data was collected, and consent was not required.

Results

The distribution of the probability of ascertainment at the county-level, in the training set, was relatively concentrated with a mean of 86.4% and a standard deviation of 3.45, along with long tails (min: 79.3%, max: 94%; Inter-quantile range: 83.8%, 88.4%). We identified 25 potential predictors of the probability of ascertainment and evaluated four sets of covariates, with the second one showing the lowest MSE of 0.049 at optimum penalization. This set included predictors with the highest explanatory power and lowest multicollinearity, including race/ethnicity, gender, age, urbanicity, poverty, schooling, migration, insurance, and wealth inequality. Variables related to income and employment were excluded due to low prediction capability. See Supplemental Materials A for the variables included in each prediction set and their corresponding MSE at optimum penalization.

Our Bayesian model showed good convergence and mixing properties. The Gelman-Rubin statistic was below 1.02 for each parameter and the multifactorial statistic was 1.03. This shows that the 5 chains overlapped and converged to the posterior distribution. Supplemental Materials B provides a visual depiction of the trace plot for each of the parameters in the final prediction model. Table 1 shows the coefficients included in the model and their mixing properties. We found that the mean intercept (2.08 95%CI: 1.79, 2.47) had the highest impact on the estimation. Higher levels of young (15-29 years) males, Non-Hispanic White, and uninsured populations are associated with a lower capacity of the surveillance system to identify HIV cases. On the contrary, higher proportions of people above the 150% federal poverty threshold, people that do not own a vehicle, and immigrants residing in the ZIP Code are associated with a greater capacity to identify HIV cases. Although the purpose of the analysis aimed at prediction and not inference, it is worth noting that none of these coefficients, except for the intercept, were significantly different from zero at 95% of confidence (Table 1).

Table 1. Variables with explanatory capacity over the probability of ascertainment, the proportion of HIV cases identified by the surveillance system. Results from the hierarchical Bayesian prediction model at the county-level for the coefficients and 95% credible intervals and good mixing properties.

Variable	Coefficient (95%CI[a])	Rhat[b]
Mean Intercept	2.08 (1.79 – 2.47)	1.02
Proportion of Male, 15-29 years of age population	-0.04 (-0.35 – 0.28)	1.00
Proportion of Non-Hispanic White population	-0.09 (-0.47 – 0.25)	1.01
Proportion of People above 150% of the federal poverty line	0.04 (-0.44 – 0.51)	1.02
Proportion of people that do not own a vehicle	0.21 (-0.24 – 0.94)	1.02
Proportion of immigrants residing in the ZIP Code	0.03 (-0.94 – 0.94)	1.01
Proportion of uninsured population	-0.11 (-0.44 – 0.22)	1.00
Gini Coefficient of inequality	0.09 (-0.31 – 0.5)	1.02
Squared of proportion of immigrants	0.09 (-0.66 – 0.94)	1.01
Proportion of people with less than high school degree squared	0.03 (-0.44 – 0.48)	1.02
Proportion of people living outside the city squared	0.1 (-0.25 – 0.49)	1.00

[a] 95%CI: 95% credible interval; Coefficients are the raw coefficients of the model output and represent a logit transformation of conditional probabilities.

[b] Gelman-Rubin statistic for good mixing of the parameters in the Bayesian model. A Rhat below 1.1 indicates adequate mixing.

Our model showed excellent prediction accuracy in the training set and both validations. We found and average prediction error of 223 in the training set, with similar differences across counties (Figure 1, panel A). The out-of-sample, county-to-county validation had an average error of 475 cases (mean of estimated cases 8,300), influenced by the difference in Fulton County, the region with the highest cases in the test set (Figure 1, panel B). The second validation, Zip Code-to-county, had a better performance with an average error of 400 cases (mean of estimated cases 8,875), due to an important improvement in the prediction for Fulton County (Figure 1, panel C). The 95%CIs of the second validation are wider than the county-to-county prediction, reflecting the greater uncertainty that comes from aggregating individual ZIP Code-data into a bigger geographic unit.

We found that the surveillance system's capacity to detect HIV cases, the probability of ascertainment, varied from 78.4% (95%CI 24.4%, 99.3%, ZIP Code: 303322) to 93.8% (95%CI 80.6%, 99.8%, ZIP Code: 30337) (Table 2). The distribution of the ascertainment probability across ZIP Codes had similar characteristics as the CDC's estimation at the county-level, with a mean of 87.1% (SD: 0.02, interquartile range: 81.7%, 92.5%). Spatially smoothed estimates produced slight differences compared to the initially predicted total, diagnosed and undiagnosed, cases because it accounts for the spatial autocorrelation of the HIV cases distribution, preventing large fluctuations across neighboring ZIP Codes. The additional parameters introduced in the model to account for the spatial structure introduce another layer of uncertainty in the estimation producing a wider 95%CI. Post smoothing, the predicted total HIV cases varied from 12 (95%CI 6, 19, ZIP Code: 30322) to 1,603 (95%CI 1,209, 1,968, ZIP Code: 30318) cases (Table 2).

Table 2. Results of the prediction of total, diagnosed and undiagnosed, HIV cases in 2018 for each ZIP Code in Atlanta, Georgia. For each ZIP Code it is reported the prediction of probability of ascertainment, total cases, and prevalence.[a]

ZIP Code	Population	Diagnosed Cases[b]	Estimated Probability of Ascertainment (95%CI)[c]	Predicted Total Cases (95%CI)[d]	Spatially Smoothed Total Cases (95%CI)[e]	Estimated Prevalence (95%CI)[f]
30322	2,025	5	0.784 (0.244 - 0.993)	6 (6 - 8)	12 (6 - 19)	544 (283 - 918)
30620	9,555	15	0.808 (0.642 - 0.920)	18 (17 - 21)	20 (12 - 31)	207 (123 - 320)
30011	11,375	24	0.809 (0.687 - 0.905)	30 (26 - 36)	31 (21 - 45)	269 (177 - 389)
30185	3,258	10	0.812 (0.481 - 0.973)	12 (11 - 13)	14 (8 - 21)	402 (227 - 641)
30115	28,314	47	0.828 (0.673 - 0.933)	53 (49 - 63)	55 (39 - 78)	194 (134 - 272)
30336	817	45	0.831 (0.510 - 0.977)	54 (48 - 64)	45 (30 - 62)	5,470 (3,671 - 7,566)
30157	35,574	90	0.833 (0.685 - 0.937)	106 (96 - 127)	112 (82 - 151)	312 (230 - 424)
30548	12,791	22	0.835 (0.654 - 0.948)	26 (23 - 43)	29 (19 - 42)	221 (142 - 325)
30187	7,240	16	0.836 (0.652 - 0.950)	19 (17 - 25)	22 (14 - 34)	302 (186 - 458)
30290	6,767	9	0.837 (0.646 - 0.953)	11 (10 - 12)	16 (9 - 26)	229 (127 - 372)
30680	29,862	86	0.840 (0.749 - 0.913)	98 (89 - 114)	98 (71 - 131)	325 (235 - 438)
30044	62,360	333	0.840 (0.620 - 0.955)	397 (371 - 435)	399 (306 - 516)	639 (489 - 827)
30144	43,494	127	0.842 (0.767 - 0.899)	141 (130 - 162)	145 (108 - 192)	331 (248 - 440)
30141	17,848	63	0.842 (0.716 - 0.935)	70 (65 - 83)	72 (52 - 100)	403 (286 - 557)
30043	65,722	278	0.842 (0.693 - 0.936)	314 (286 - 362)	311 (237 - 402)	473 (360 - 611)
30047	48,671	183	0.844 (0.673 - 0.944)	206 (188 - 244)	221 (169 - 299)	453 (345 - 613)
30228	30,250	206	0.844 (0.730 - 0.932)	242 (219 - 284)	242 (184 - 316)	800 (606 - 1,044)
30116	18,726	47	0.844 (0.725 - 0.925)	54 (48 - 65)	57 (40 - 79)	301 (210 - 421)
30052	47,059	136	0.846 (0.766 - 0.914)	159 (147 - 177)	167 (126 - 225)	355 (267 - 477)
30273	11,686	122	0.846 (0.686 - 0.946)	133 (123 - 158)	135 (100 - 181)	1,154 (854 - 1,542)
30019	31,148	76	0.846 (0.749 - 0.923)	86 (79 - 102)	91 (67 - 123)	290 (213 - 393)
30101	42,914	109	0.847 (0.743 - 0.924)	122 (110 - 154)	126 (94 - 169)	293 (217 - 393)
30102	30,259	82	0.848 (0.788 - 0.902)	89 (83 - 116)	91 (66 - 124)	299 (215 - 408)
30094	25,969	114	0.849 (0.719 - 0.947)	131 (119 - 149)	138 (103 - 188)	530 (393 - 721)
30135	49,620	261	0.849 (0.745 - 0.932)	292 (267 - 348)	310 (239 - 414)	623 (481 - 832)
30188	42,242	98	0.850 (0.752 - 0.926)	114 (106 - 125)	116 (86 - 155)	273 (201 - 366)
30518	35,103	86	0.850 (0.742 - 0.924)	110 (89 - 179)	109 (79 - 145)	309 (225 - 413)
30122	17,380	126	0.851 (0.747 - 0.929)	146 (137 - 158)	151 (113 - 202)	867 (650 - 1,158)
30180	26,901	92	0.851 (0.730 - 0.938)	110 (99 - 127)	111 (81 - 149)	411 (301 - 551)
30084	29,498	323	0.851 (0.708 - 0.940)	356 (327 - 440)	364 (281 - 475)	1,234 (952 - 1,610)
30296	20,453	289	0.851 (0.703 - 0.949)	350 (305 - 442)	348 (264 - 450)	1,698 (1,288 - 2,198)
30066	44,095	112	0.853 (0.761 - 0.924)	130 (121 - 142)	135 (101 - 182)	306 (227 - 411)
30152	32,736	73	0.853 (0.758 - 0.924)	83 (77 - 92)	90 (66 - 124)	274 (200 - 377)
30288	7,113	85	0.854 (0.711 - 0.944)	96 (87 - 119)	104 (76 - 143)	1,452 (1,066 - 1,998)
30360	11,499	99	0.854 (0.677 - 0.952)	110 (102 - 121)	113 (83 - 151)	976 (721 - 1,308)
30062	50,752	135	0.854 (0.750 - 0.928)	147 (137 - 171)	156 (118 - 210)	306 (231 - 412)

30238	26,958	327	0.854 (0.733 - 0.943)	391 (359 - 437)	380 (288 - 486)	1,409 (1,066 - 1,802)
30274	25,113	390	0.855 (0.705 - 0.953)	449 (408 - 518)	451 (344 - 586)	1,793 (1,367 - 2,333)
30040	42,515	71	0.855 (0.744 - 0.931)	84 (77 - 96)	86 (62 - 116)	200 (145 - 273)
30045	26,378	158	0.855 (0.764 - 0.925)	185 (170 - 207)	180 (135 - 234)	682 (508 - 885)
30068	26,214	54	0.855 (0.707 - 0.951)	62 (55 - 88)	71 (51 - 100)	268 (191 - 380)
30041	40,910	45	0.855 (0.740 - 0.935)	52 (49 - 59)	57 (40 - 80)	137 (95 - 193)
30223	29,873	141	0.856 (0.727 - 0.943)	167 (157 - 179)	169 (125 - 225)	563 (418 - 751)
30024	49,646	70	0.857 (0.734 - 0.936)	81 (74 - 96)	87 (64 - 120)	175 (127 - 241)
30134	33,990	190	0.857 (0.781 - 0.923)	224 (205 - 255)	220 (165 - 284)	645 (485 - 833)
30127	48,410	213	0.857 (0.767 - 0.929)	267 (218 - 418)	268 (204 - 347)	552 (421 - 715)
30004	41,905	88	0.858 (0.674 - 0.953)	99 (92 - 113)	103 (76 - 138)	244 (180 - 328)
30008	23,733	206	0.858 (0.712 - 0.946)	242 (221 - 279)	239 (180 - 311)	1,006 (754 - 1,310)
30064	38,014	111	0.859 (0.727 - 0.944)	129 (116 - 159)	139 (104 - 191)	365 (273 - 500)
30297	21,029	270	0.860 (0.640 - 0.967)	318 (287 - 369)	330 (254 - 435)	1,566 (1,205 - 2,068)
30017	16,202	58	0.860 (0.762 - 0.934)	71 (62 - 85)	73 (52 - 100)	446 (316 - 613)
30022	51,531	101	0.860 (0.710 - 0.946)	115 (106 - 131)	126 (94 - 173)	244 (182 - 335)
30517	9,402	22	0.860 (0.740 - 0.937)	25 (23 - 28)	26 (16 - 39)	272 (170 - 406)
30260	18,756	208	0.861 (0.713 - 0.947)	246 (219 - 308)	248 (187 - 324)	1,318 (997 - 1,727)
30078	27,153	82	0.861 (0.770 - 0.926)	97 (89 - 108)	103 (76 - 140)	378 (278 - 515)
30294	31,907	410	0.861 (0.718 - 0.962)	465 (425 - 543)	472 (365 - 613)	1,479 (1,141 - 1,921)
30281	53,251	344	0.861 (0.787 - 0.929)	405 (361 - 511)	422 (327 - 559)	792 (613 - 1,048)
30012	21,806	128	0.861 (0.797 - 0.917)	144 (130 - 173)	149 (111 - 199)	680 (508 - 910)
30519	31,440	83	0.861 (0.782 - 0.923)	98 (89 - 114)	98 (72 - 131)	310 (227 - 414)
30046	26,803	171	0.862 (0.692 - 0.954)	201 (182 - 236)	198 (148 - 258)	737 (549 - 962)
30214	24,544	94	0.862 (0.778 - 0.928)	108 (97 - 144)	119 (88 - 163)	482 (359 - 664)
30039	32,759	209	0.862 (0.756 - 0.938)	246 (217 - 327)	246 (187 - 321)	751 (568 - 979)
30236	37,594	353	0.863 (0.758 - 0.938)	427 (372 - 542)	438 (338 - 572)	1,163 (898 - 1,519)
30291	16,189	288	0.865 (0.731 - 0.962)	341 (307 - 408)	336 (254 - 436)	2,075 (1,563 - 2,691)
30215	27,869	68	0.865 (0.750 - 0.945)	79 (72 - 91)	87 (63 - 122)	312 (226 - 437)
30096	50,607	292	0.867 (0.661 - 0.966)	333 (303 - 382)	336 (257 - 438)	663 (506 - 864)
30313	8,986	217	0.867 (0.470 - 0.996)	249 (230 - 276)	263 (201 - 353)	2,923 (2,233 - 3,927)
30082	21,068	150	0.867 (0.799 - 0.924)	163 (152 - 186)	173 (130 - 233)	817 (616 - 1,105)
30106	16,327	144	0.868 (0.792 - 0.933)	175 (145 - 307)	173 (130 - 224)	1,055 (792 - 1,372)
30067	37,132	378	0.869 (0.702 - 0.958)	437 (383 - 606)	425 (322 - 543)	1,144 (867 - 1,461)
30168	18,630	223	0.869 (0.715 - 0.959)	255 (232 - 296)	259 (196 - 341)	1,386 (1,049 - 1,827)
30263	42,899	148	0.870 (0.803 - 0.926)	170 (154 - 208)	173 (129 - 231)	402 (300 - 536)
30316	26,540	681	0.870 (0.792 - 0.939)	753 (685 - 979)	747 (576 - 956)	2,815 (2,168 - 3,598)
30092	25,374	238	0.871 (0.774 - 0.937)	279 (258 - 305)	269 (202 - 344)	1,058 (794 - 1,355)
30087	30,536	291	0.871 (0.793 - 0.937)	345 (319 - 380)	351 (269 - 459)	1,147 (878 - 1,501)
30060	27,193	338	0.872 (0.676 - 0.974)	392 (348 - 500)	359 (267 - 447)	1,318 (979 - 1,643)
30075	43,221	102	0.873 (0.753 - 0.957)	117 (106 - 134)	124 (92 - 167)	285 (212 - 384)
30033	26,667	288	0.873 (0.735 - 0.954)	309 (289 - 358)	321 (248 - 424)	1,203 (927 - 1,589)
30213	22,991	286	0.875 (0.787 - 0.947)	331 (299 - 400)	324 (246 - 414)	1,405 (1,066 - 1,799)
30306	19,383	320	0.877 (0.681 - 0.977)	400 (348 - 499)	412 (317 - 540)	2,121 (1,635 - 2,783)
30038	29,768	448	0.877 (0.766 - 0.961)	519 (453 - 743)	510 (389 - 651)	1,710 (1,306 - 2,185)
30005	26,601	54	0.877 (0.713 - 0.964)	64 (58 - 73)	65 (46 - 88)	242 (173 - 330)
30088	20,776	386	0.879 (0.753 - 0.965)	424 (391 - 509)	426 (325 - 554)	2,047 (1,560 - 2,665)
30093	40,020	407	0.879 (0.624 - 0.988)	444 (411 - 540)	439 (335 - 565)	1,096 (836 - 1,410)

30097	33,645	72	0.880 (0.655 - 0.978)	79 (73 - 97)	85 (62 - 117)	252 (183 - 348)
30268	7,321	47	0.880 (0.828 - 0.928)	55 (51 - 59)	54 (38 - 74)	728 (506 - 1,004)
30303	5,802	374	0.881 (0.604 - 0.989)	435 (399 - 494)	425 (322 - 544)	7,316 (5,537 - 9,370)
30126	29,079	246	0.882 (0.783 - 0.948)	288 (258 - 356)	307 (236 - 411)	1,052 (811 - 1,412)
30071	17,776	147	0.882 (0.763 - 0.963)	167 (151 - 198)	170 (127 - 225)	953 (714 - 1,263)
30349	53,049	1314	0.883 (0.766 - 0.965)	1,569 (1,404 - 1,895)	1,529 (1,178 - 1,925)	2,882 (2,219 - 3,628)
30034	35,492	751	0.883 (0.759 - 0.969)	896 (804 - 1,050)	896 (689 - 1,155)	2,523 (1,939 - 3,253)
30058	42,033	604	0.885 (0.779 - 0.965)	660 (615 - 760)	657 (504 - 845)	1,563 (1,199 - 2,010)
30035	16,431	406	0.886 (0.745 - 0.974)	481 (436 - 546)	474 (361 - 607)	2,879 (2,193 - 3,691)
30319	33,090	226	0.886 (0.773 - 0.958)	261 (241 - 286)	282 (218 - 380)	851 (658 - 1,148)
30076	34,194	159	0.886 (0.756 - 0.966)	185 (173 - 204)	182 (136 - 238)	532 (397 - 695)
30327	18,566	88	0.887 (0.618 - 0.991)	105 (96 - 122)	116 (86 - 161)	622 (460 - 863)
30080	41,276	369	0.889 (0.794 - 0.953)	434 (400 - 485)	446 (344 - 585)	1,079 (831 - 1,417)
30340	23,411	313	0.889 (0.734 - 0.980)	370 (330 - 446)	365 (277 - 471)	1,559 (1,182 - 2,011)
30079	2,303	46	0.890 (0.790 - 0.962)	56 (49 - 75)	56 (39 - 76)	2,401 (1,684 - 3,298)
30083	40,539	889	0.891 (0.751 - 0.970)	1,044 (958 - 1,186)	1,030 (793 - 1,310)	2,539 (1,955 - 3,230)
30345	19,229	314	0.891 (0.769 - 0.972)	366 (327 - 430)	361 (274 - 466)	1,877 (1,423 - 2,423)
30305	20,213	190	0.892 (0.634 - 0.990)	219 (203 - 240)	233 (178 - 315)	1,152 (880 - 1,554)
30338	27,250	106	0.893 (0.762 - 0.966)	121 (115 - 129)	135 (101 - 187)	493 (368 - 683)
30341	26,593	359	0.893 (0.751 - 0.974)	434 (377 - 556)	425 (323 - 544)	1,597 (1,213 - 2,046)
30009	11,321	36	0.894 (0.781 - 0.965)	41 (38 - 46)	41 (28 - 58)	361 (243 - 511)
30030	22,190	227	0.895 (0.795 - 0.962)	267 (246 - 300)	282 (216 - 377)	1,267 (972 - 1,697)
30032	39,653	1096	0.895 (0.750 - 0.979)	1,299 (1,186 - 1,475)	1,265 (971 - 1,594)	3,188 (2,448 - 4,020)
30354	11,767	341	0.899 (0.742 - 0.988)	532 (344 - 1,397)	517 (392 - 657)	4,386 (3,327 - 5,580)
30350	29,308	381	0.899 (0.751 - 0.978)	449 (401 - 540)	412 (305 - 513)	1,403 (1,040 - 1,748)
30307	15,228	194	0.900 (0.719 - 0.984)	226 (207 - 257)	233 (177 - 307)	1,524 (1,161 - 2,013)
30339	16,535	209	0.901 (0.794 - 0.968)	249 (221 - 305)	247 (187 - 322)	1,493 (1,125 - 1,943)
30329	23,563	459	0.904 (0.760 - 0.982)	529 (470 - 675)	518 (394 - 662)	2,196 (1,672 - 2,806)
30314	19,238	949	0.906 (0.710 - 0.992)	1,042 (963 - 1,215)	1,033 (793 - 1,324)	5,367 (4,117 - 6,881)
30344	25,758	833	0.907 (0.781 - 0.985)	974 (866 - 1,160)	976 (754 - 1,254)	3,787 (2,925 - 4,867)
30317	10,172	207	0.908 (0.807 - 0.973)	236 (210 - 327)	237 (178 - 311)	2,323 (1,750 - 3,048)
30021	17,016	350	0.909 (0.517 - 0.999)	419 (371 - 521)	416 (317 - 537)	2,443 (1,859 - 3,156)
30331	43,649	914	0.909 (0.788 - 0.981)	1,076 (960 - 1,297)	1,073 (829 - 1,372)	2,457 (1,899 - 3,143)
30324	21,720	960	0.910 (0.823 - 0.977)	1,061 (987 - 1,190)	1,000 (752 - 1,248)	4,601 (3,459 - 5,744)
30309	20,415	759	0.913 (0.735 - 0.990)	883 (819 - 976)	860 (656 - 1,091)	4,210 (3,209 - 5,341)
30002	4,783	88	0.914 (0.840 - 0.977)	99 (92 - 116)	101 (74 - 136)	2,102 (1,533 - 2,840)
30318	41,426	1592	0.915 (0.782 - 0.986)	1,730 (1,605 - 2,059)	1,603 (1,209 - 1,968)	3,868 (2,917 - 4,749)
30363	2,573	32	0.915 (0.696 - 0.995)	38 (35 - 42)	42 (28 - 60)	1,597 (1,061 - 2,319)
30342	23,898	206	0.915 (0.759 - 0.988)	241 (218 - 289)	245 (186 - 322)	1,023 (777 - 1,344)
30328	25,905	179	0.918 (0.795 - 0.983)	223 (198 - 261)	225 (172 - 294)	868 (661 - 1,133)
30308	14,473	851	0.923 (0.777 - 0.993)	998 (909 - 1,160)	963 (731 - 1,215)	6,648 (5,051 - 8,389)
30346	4,259	50	0.923 (0.755 - 0.992)	60 (55 - 67)	58 (41 - 80)	1,359 (948 - 1,867)
30326	4,496	69	0.924 (0.791 - 0.992)	76 (71 - 83)	77 (55 - 104)	1,694 (1,211 - 2,307)
30312	16,571	634	0.924 (0.809 - 0.989)	702 (644 - 812)	698 (537 - 896)	4,212 (3,238 - 5,402)
30315	26,881	1004	0.925 (0.773 - 0.992)	1,190 (1,061 - 1,397)	1,175 (902 - 1,499)	4,368 (3,355 - 5,573)
30311	26,100	789	0.928 (0.744 - 0.997)	932 (852 - 1,050)	942 (727 - 1,222)	3,607 (2,783 - 4,679)
30310	22,088	967	0.930 (0.711 - 0.998)	1,100 (1,021 - 1,235)	1,095 (845 - 1,402)	4,955 (3,822 - 6,344)
30337	9,185	284	0.938 (0.806 - 0.998)	331 (310 - 357)	330 (250 - 429)	3,587 (2,716 - 4,665)

[a] Data is sorted increasingly by the probability of ascertainment. The Probability of Ascertainment was approximated to 3-decimal digits. Total cases approximated to the nearest unit.

[b] Diagnosed cases: HIV cases identified by the surveillance system in each ZIP Code in Atlanta Georgia in 2018. Data collected from Emory University's data visualization tool AIDSVu.

[c] Probability of Ascertainment: Passive surveillance system's capacity to diagnose and report an HIV case at the ZIP Code-level and 95% credible interval (95%CI). Estimated using the results of the hierarchical Bayesian prediction model at the county-level coupled with ZIP Code level data of sociodemographic determinants of HIV spreading collected from the American Community Survey (ACS).

[d] Predicted Total cases: Diagnosed and undiagnosed cases of HIV. Estimated using the predictive posterior distribution of the coefficients of the probability of ascertainment applied to Zip Code-level data of sociodemographic determinants of HIV spreading and diagnosed HIV cases collected from AIDSVu.

[e] Spatially Smoothed Total Cases: Predicted Total Cases after applying the spatially smoothing model to account for the spatial autocorrelation of the HIV cases.

[f] Prevalence: Cases per 100,000 people. Estimated using the spatially smoothed total cases.

We found that the prevalence of HIV presented only slight differences after the addition of undiagnosed cases in several areas of Atlanta (Figure 2, panel A). Nonetheless, there are noticeable differences in the outer areas of the city, specifically the north and west regions. This is more easily observed in the geographical distribution of the probability of ascertainment, where distance from the city-center is negatively correlated to the proportion of HIV cases captured by the surveillance system (Figure 2, panel B). Furthermore, the uncertainty in the estimation varies across ZIP Codes and increases relative to the closeness to the city-center: the further from the city-center the higher the uncertainty, particularly for ZIP Codes located in the southwest area (Figure 2, panel C).

Discussion

We used a hierarchical Bayesian model to predict the number of total, diagnosed and undiagnosed, HIV cases at the ZIP Code-level in Atlanta using passive surveillance data and social determinants of HIV spreading as data inputs. Our model showed good mixing properties and excellent predictive accuracy in the training set and both validations, county-to-county and ZIP code-to-county. The use of Bayesian statistics in our analytical approach allowed us to incorporate the complexities of the data generation process in the prediction model, as well as the main sources of uncertainty to obtain reliable estimates.[33] Furthermore, we fitted the data to a spatial statistics model to account for the spatial autocorrelation of the data and obtain internally consistent estimates throughout all ZIP Codes within the city.

On average, our prediction efforts were 475 cases further from the CDC estimates in the county-to-county validation using the four Georgia counties included in the EHE. This prediction error was 252 cases greater than the same metric in the training set. There are two main reasons for such a difference in accuracy. First, the prediction was entirely out-of-sample, which means that the training set did not use any information of the Georgia counties to estimate the coefficients, and therefore higher errors are expected.[34] Second, the validation set had fewer observations, so the errors of every county had a greater weight on the MAE. This is particularly evident for Fulton County whose performance improvement in the ZIP Code-to-county validation led to a reduction of 75 cases in the overall MAE, compared to the county-to-county validation. The prediction on all sets generated narrower uncertainty bounds compared to the CDC-reported estimates. This is likely a byproduct the low variability observed in the probability of ascertainment in the train set, which ranges from 79.3% to 94%, and the high-accuracy of the predictors of the probability of ascertainment (Table 1).

These estimates highlight significant differences in the capacity of the surveillance system to identify HIV cases varies across ZIP Codes, varying from 78.4% to 93.8% (Table 2). Rather than differential efforts made by the local departments of health to detect HIV cases, we believe the estimated variability is a consequence of the number of cases arising from each ZIP Code (in areas with fewer cases, detecting each additional one demands greater efforts), and differences in key social determinants of health which influence healthcare-seeking behavior and access to healthcare services.[24,35] This implies that the same effort to identify HIV cases through the promotion of diagnosis and awareness would yield varying results depending on the ZIP Code and its demographic characteristics. To achieve equality in the diagnosis of HIV within a city, it is important to recognize that certain communities have a greater concentration of people not seeking care in traditional settings (i.e., hospitals and clinics), due to lack of insurance, poverty, housing insecurity, limited availability of services, stigma, and discrimination.[36,37] Therefore, these communities require greater outreach efforts. Our results suggest that the ZIP Codes farther from the city-center are more prone to depict lower ascertainment probabilities, which echoes the findings from studies on the impact of rurality on HIV testing outcomes.[38]

Another implication of our findings is the effect on the assessment of the relative burden of HIV across ZIP Codes once undiagnosed cases are accounted for in the estimation of prevalence. Adding undiagnosed cases allows for a more realistic identification of high-risk and underserved areas and permits a more efficient allocation of resources.[39,40] While this is relevant for a citywide assessment, we believe the most benefits would be extracted at the county-level given that the organizational structure of the Local Health Departments in Atlanta and the Georgia Department of Public Health follows a shared governance model.[41] This is particularly important for counties whose ZIP codes have relatively low prevalence of HIV, such as Cobb and Gwinnett, where the addition of even a few undiagnosed cases could reveal a different distribution of risk across areas, suggesting greater needs for HIV-related services than previously thought. Under this governance model, local departments have some autonomy to make decisions on delivery of health services and allocations of resources, and effectively represent the first line of protection against public health threats.[42] Thus, our results could provide important evidence of the true burden of disease within counties and help inform programmatic activities to increase outreach and enhance delivery services in underserved areas.

This study has limitations. First, the capacity of the HIV surveillance system to identify cases across different jurisdictions is likely influenced by several factors, including collection and storage of data capacities, testing and diagnosis infrastructure, individual’s health care seeking behavior, determinants of access to care, population growth and migration patterns, as well as the legal and policy framework.[43,44] Of these, we focused on social determinants of health in our analysis, which fairly represented healthcare care seeking behavior and demographic characteristics. However, we lacked information mainly on the supply side of healthcare resources (e.g., healthcare resources used to provide HIV-related services such as facilitates, pharmacies, and personnel), and determinants of the disease status (e.g., proportion of men who have sex with men, intravenous drug users, sexual workers). Increasing the diversity of prediction variables could improve the accuracy of the ascertainment probability estimates and provide greater insight in the determinants of the surveillance system’s ascertainment capacity. Second, given that the main purpose of this analysis was to provide a path to estimate undiagnosed cases at the ZIP Code level, which is otherwise unavailable, we were not able to conduct a ZIP Code to ZIP Code validation. The second validation, ZIP Code-to-county, provides a contrast for the ZIP Code-aggregated results, but it is not a complete assessment at the ZIP Code level because the conclusions are drawn for counties. Directly validating ZIP Code data would only be possible if ZIP Code level undiagnosed cases are estimated using alternative statistical modeling approaches, for example by leveraging surveillance laboratory data, or by directly conducting prevalence studies with such geographic granularity built-in.[45,46]

Conclusions

This study's main conclusions are twofold. First, given how heterogeneous the ascertainment probability is at a ZIP Code level, correcting for undiagnosed cases is not a trivial process and has the potential to change the relative burden of disease within a city and more importantly in a county. Our study described a robust statistical approach, solely based on publicly available data, which can effectively complement the information obtained from passive surveillance, especially when more resource-intensive approaches (such as laboratory-based estimations) are not available or are unfeasible to implement.

Second, our results suggest that the public health system's case identification capacity diminishes as ZIP Codes move further from the city center. This is not surprising given that this is a passive system that relies on potentially infected people to seek diagnostic services, and the availability of such services reduces the further away from urban centers. Future resource allocation decisions should consider the need for a greater presence of the public health system in those areas to increase the likelihood of identifying HIV cases and referring them to treatment.

Competing interest

The authors have no conflicts of interest to declare.

Acknowledgements

EMS and AB designed the study and methodological framework. EMS conducted the analysis and prepared the initial manuscript draft. AB supervised the execution of the analysis and reviewed the manuscript. All authors approved the submitted version of the manuscript.

Data Availability Statement

Data sharing is not applicable to this article as no new data was created in this study. Secondary analysis of publicly available information.

References

1. Centers for Disease Control and Prevention. HIV Surveillance - United States, 1981--2008. MMWR Morb Mortal Wkly Rep. 2011;60(21):689-693. PMID: 21637182.

2. McCree DH, Young SR, Henny KD, Cheever L, McCray E. U.S. Centers for Disease Control and Prevention and Health Resources and Services Administration Initiatives to Address Disparate Rates of HIV Infection in the South. AIDS Behav. 2019 Oct 1;23(3):313-318. PMID: 31321635.

3. Centers for Disease Control and Prevention. HIV Surveillance Report, 2018 (Updated) [Internet]. 2020 May [cited 2023 Nov 18] p. 119. Report No.: vol 31. Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-report-2018-updated-vol-31.pdf

4. Panagiotoglou D, Olding M, Enns B, Feaster D, del Rio C, Metsch L, et al. Building the case for localized approaches to HIV: structural conditions and health system capacity to address the HIV/AIDS epidemic in six US cities. AIDS Behav. 2018 Sep;22(9):3071-3082. PMID: 29802550.

5. El-Sadr WM, Mayer KH, Rabkin M, Hodder SL. AIDS in America — Back in the Headlines at Long Last. N Engl J Med. 2019 May 23;380(21):1985-1987. PMID: 31042822.

6. Guilamo-Ramos V, Thimm-Kaiser M, Benzekri A, Chacón G, López OR, Scaccabarrozzi L, et al. The Invisible US Hispanic/Latino HIV Crisis: Addressing Gaps in the National Response. Am J Public Health. 2019 Nov 14;110(1):27-31. PMID: 31725313.

7. Reif S, Safley D, McAllaster C, Wilson E, Whetten K. State of HIV in the US Deep South. J Community Health. 2017 Oct;42(5):844-853. PMID: 28247067.

8. Ransome Y, Kawachi I, Braunstein S, Nash D. Structural inequalities drive late HIV diagnosis: The role of black racial concentration, income inequality, socioeconomic deprivation, and HIV testing. Health Place. 2016 Nov;42:148-158. PMID: 27770671.

9. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV Epidemic: A Plan for the United States. JAMA. 2019 Mar 5;321(9):844-845. PMID: 30730529.

10. Hall HI, Brooks JT, Mermin J. Can the United States achieve 90–90–90?: Curr Opin HIV AIDS. 2019 Nov;14(6):464-470. PMID: 31425180.

11. Hernández-Romieu AC, Sullivan PS, Rothenberg R, Grey J, Luisi N, Kelley CF, et al. Heterogeneity of HIV prevalence among the sexual networks of Black and White MSM in Atlanta: illuminating a mechanism for increased HIV risk for young Black MSM. Sex Transm Dis. 2015 Sep;42(9):505-512. PMID: 26267877.

12. Schleimer JP, Buggs SA, McCort CD, Pear VA, Biasi AD, Tomsich E, et al. Neighborhood Racial and Economic Segregation and Disparities in Violence During the COVID-19 Pandemic. Am J Public Health. 2022 Jan;112(1):144-153. PMID: 34882429.

13. Eberth JM, Hung P, Benavidez GA, Probst JC, Zahnd WE, McNatt MK, et al. The Problem Of The Color Line: Spatial Access To Hospital Services For Minoritized Racial And Ethnic Groups. Health Aff Proj Hope. 2022 Feb;41(2):237-246. PMID: 35130071.

14. Kay ES, Batey DS, Mugavero MJ. The HIV treatment cascade and care continuum: updates, goals, and recommendations for the future. AIDS Res Ther. 2016 Nov 8;13:35. PMID: 27826353.

15. Richard M. Selik, Eve D. Mokotoff, Bernard Branson, S. Michelle Owen, Suzanne Whitmore, H. Irene Hall. Revised Surveillance Case Definition for HIV Infection — United States, 2014. MMWR Morb Mortal Wkly Rep. 2014;63(RR03):1-10. PMID: 24717910.

16. Sullivan PS, Woodyatt C, Koski C, Pembleton E, McGuinness P, Taussig J, et al. A Data Visualization and Dissemination Resource to Support HIV Prevention and Care at the Local Level: Analysis and Uses of the AIDSVu Public Data Resource. J Med Internet Res. 2020 Oct 23;22(10):e23173. PMID: 33095177.

17. Sullivan PS, McKenna MT, Waller LA, Williamson GD, Lee LM. Analyzing and Interpreting Public Health Surveillance Data. In: Lee LM, Teutsch SM, Thacker SB, St. Louis ME, editors. Principles & Practice of Public Health Surveillance. Oxford University Press; 2010.

18. Souty C, Turbelin C, Blanchon T, Hanslik T, Le Strat Y, Boëlle PY. Improving disease incidence estimates in primary care surveillance systems. Popul Health Metr. 2014 Jul 26;12(1):19. PMID: 25435814.

19. Song R, Hall HI, Green TA, Szwarcwald CL, Pantazis N. Using CD4 Data to Estimate HIV Incidence, Prevalence, and Percent of Undiagnosed Infections in the United States. J Acquir Immune Defic Syndr 1999. 2017 01;74(1):3-9. PMID: 27509244.

20. Centers for Disease Control and Prevention. Estimated HIV incidence and prevalence in the United States, 2014–2018. HIV Surveillance Supplemental Report 2020 [Internet]. 2020 May [cited 2023 Dec 1] p. 78. Report No.: 25(No. 1). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-25-1.pdf

21. US Preventive Services Task Force, Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, et al. Screening for HIV Infection: US Preventive Services Task Force Recommendation Statement. JAMA. 2019 Jun 18;321(23):2326-2336. PMID: 31184701.

22. Hall HI, Song R, Szwarcwald CL, Green T. Brief Report: Time From Infection With the Human Immunodeficiency Virus to Diagnosis, United States. JAIDS J Acquir Immune Defic Syndr. 2015 Jun 1;69(2):248-251. PMID: 25714245.

23. Sullivan PS, Peterson J, Rosenberg ES, Kelley CF, Cooper H, Vaughan A, et al. Understanding Racial HIV/STI Disparities in Black and White Men Who Have Sex with Men: A Multilevel Approach. PLoS ONE. 2014 Mar 7;9(3):e90514. PMID: 24608176.

24. Centers for Disease Control and Prevention. Social Determinants of Health Among Adults with Diagnosed HIV Infection, 2018. HIV Surveillance Supplemental Report 2020 [Internet]. 2020 Nov [cited 2024 Jan 10] p. 78. Report No.: 25(No. 3). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-supplemental-report-2020-vol25-no3.pdf

25. Hess KL, Hu X, Lansky A, Mermin J, Hall HI. Lifetime risk of a diagnosis of HIV infection in the United States. Ann Epidemiol. 2017 Apr;27(4):238-243. PMID: 28325538.

26. van Zwet E. A default prior for regression coefficients. Stat Methods Med Res. 2019 Dec;28(12):3799-3807. PMID: 30543154.

27. Boonstra PS, Barbaro RP, Sen A. Default Priors for the Intercept Parameter in Logistic Regressions. Comput Stat Data Anal. 2019 May;133:245-256. PMID: 31530966.

28. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci.

1992 Nov;7(4):457–72.

29. Martyn Plummer. JAGS Version 4.3.0 user manual [Internet]. 2017 Jun [cited 2021 May 15]. Available from: https://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf

30. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991 Mar 1;43(1):1–20.

31. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B Stat Methodol. 2009;71(2):319–92.

32. HUD Office of Policy Development and Research. HUD User. [cited 2022 Mar 16]. HUD USPS ZIP Code Crosswalk Files. Available from: https://www.huduser.gov/portal/datasets/usps_crosswalk.html

33. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. Int J Epidemiol. 2007 Feb;36(1):195-202. PMID: 17329317.

34. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. J Anal Test. 2018 Jul 1;2(3):249-262. PMID: 30842888.

35. Mahajan AP, Sayles JN, Patel VA, Remien RH, Ortiz D, Szekeres G, et al. Stigma in the HIV/AIDS epidemic: A review of the literature and recommendations for the way forward. AIDS Lond Engl. 2008 Aug;22(Suppl 2):S67-S79. PMID: 18641472.

36. Valdiserri RO. Improving Outcomes Along the HIV Care Continuum: Paying Careful Attention to the Non-Biologic Determinants of Health. Public Health Rep. 2014;129(4):319-321. PMID: 24982533.

37. Menza TW, Hixson LK, Lipira L, Drach L. Social Determinants of Health and Care Outcomes Among People With HIV in the United States. Open Forum Infect Dis. 2021 Jun 22;8(7):ofab330. PMID: 34307729.

38. Tran L, Tran P, Tran L. Influence of Rurality on HIV Testing Practices Across the United States, 2012-2017. AIDS Behav. 2020 Feb;24(2):404-417. PMID: 30762188.

39. Alistar SS, Brandeau ML. Decision making for HIV prevention and treatment scale up: Bridging the gap between theory and practice. Med Decis Making. 2012 Jan;32(1):105-117. PMID: 21191118.

40. Drake TL, Lubell Y, Kyaw SS, Devine A, Kyaw MP, Day NPJ, et al. Geographic Resource Allocation Based on Cost Effectiveness: An Application to Malaria Policy. Appl Health Econ Health Policy. 2017;15(3):299-306. PMID: 28185133.

41. NACCHO. 2022 National Profile of Local Health Departments [Internet]. Washington, D.C.: National Association of County & City Health Officials; 2022 [cited 2024 Apr 10]. Available from: https://www.naccho.org/uploads/downloadable-resources/NACCHO_2022_Profile_Report.pdf

42. ASTHO. ASTHO Profile of State Public Health volume 4 [Internet]. Washington, D.C.: Association of State and Territorial Health Officials; 2017 [cited 2024 Apr 10]. Available from: https://www.astho.org/globalassets/pdf/profile/profile-stph-vol-4.pdf

43. Schmidt MA, Mokotoff ED. HIV/AIDS surveillance and prevention: improving the

characterization of HIV transmission. Public Health Rep. 2003;118(3):197-204. PMID: 12766214.

44. CohenN SM, Gray KM, Ocfemia MCB, Johnson AS, Hall HI. The Status of the National HIV Surveillance System, United States, 2013. Public Health Rep. 2014;129(4):335-341. PMID: 24982536.

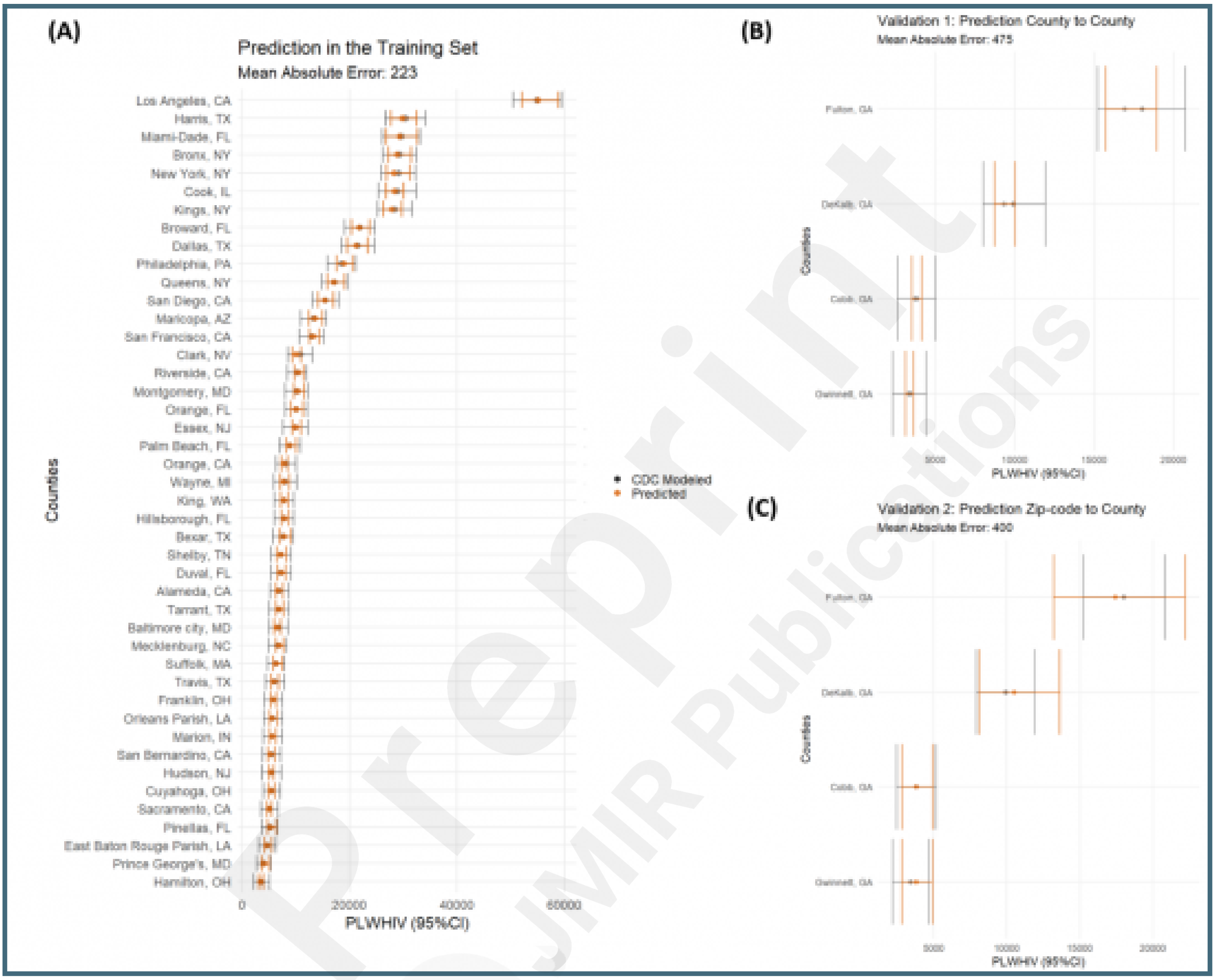
45. Working Group on Estimation of HIV Prevalence in Europe. HIV in hiding: methods and data requirements for the estimation of the number of people living with undiagnosed HIV. AIDS Lond Engl. 2011 May 15;25(8):1017-1023. PMID: 21422986.

46. Brookmeyer R. Measuring the HIV/AIDS epidemic: approaches and challenges. Epidemiol Rev. 2010;32:26-37. PMID: 20203104.

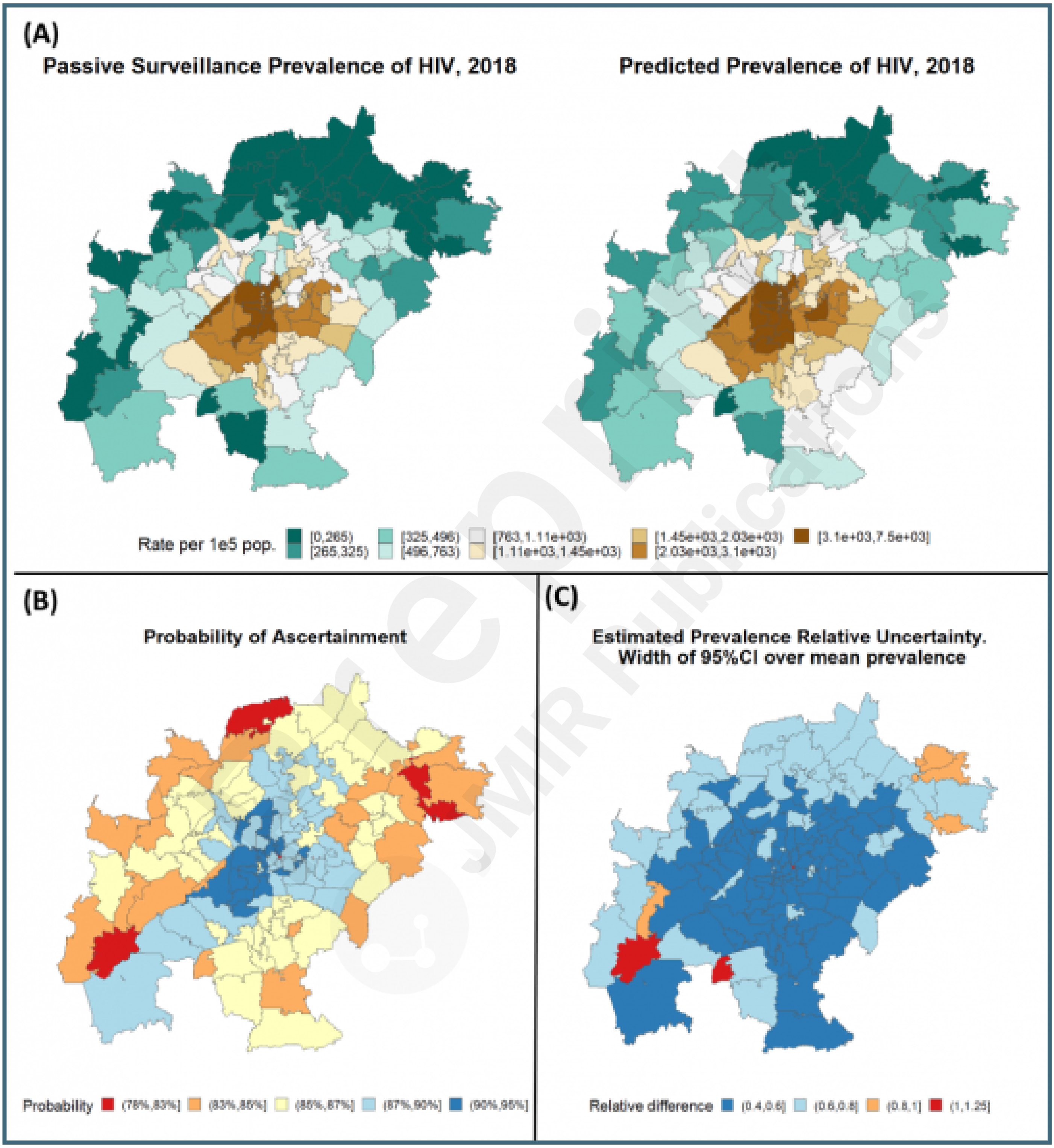
Supplementary Files

Figures

Accuracy assessment of the Bayesian hierarchical model for the prediction of total, diagnosed and undiagnosed, HIV cases in 2018 in the (A) forty-four counties included in the training set, the (B) four counties in the conunty-to-county validations, and at the (C) ZIP Code level prediction aggregated at the county level. Key: PLWHI: People Living with HIV; 95%CI: 95% credible intervals.



Estimated Prevalence of HIV in 2018 at the ZIP Code-level. (A) Comparing passive surveillance data (diagnoses cases only) and predicted (total cases), (B) Predicted proportion of cases identified through passive surveillance, and (C) Measure of uncertainty in the prediction of undiagnosed cases. Key: 95%CI: 95% credible intervals.



Multimedia Appendixes

Variables included in the feature selection algorithm for the prediction model.
URL: http://asset.jmir.pub/assets/a6db71b604522e674765604e5dff2067.docx

Coefficients’ mixing performance.
URL: http://asset.jmir.pub/assets/a5e6859ed7d019a336a8bec0862971d6.docx