Received: 11 November 2022 | Revised: 31 January 2023 | Accepted: 11 February 2023

DOI: 10.1002/sta4.555

ORIGINAL ARTICLE

WILEY

Bias correction for nonignorable missing counts of areal HIV new diagnosis

Tianyi Qu[1] | Bo Li[1] | Man-pui Sally Chan[2] | Dolores Albarracin[2]

[1]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, 61820, USA

[2]Annenberg School for Communication, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA

Correspondence
Tianyi Qu, Department of Statistics, University of Illinois at Urbana-Champaign, 605 E. Springfield Ave., Champaign, Illinois 61820, USA.
Email: tianyiq3@illinois.edu

Funding information
National Institute of Allergy and Infectious Diseases; National Institute of Mental Health; National Institute on Drug Abuse

Public health data, such as HIV new diagnoses, are often left-censored due to confidentiality issues. Standard analysis approaches that assume censored values as missing at random often lead to biased estimates and inferior predictions. Motivated by the Philadelphia areal counts of HIV new diagnosis for which all values less than or equal to 5 are suppressed, we propose two methods to reduce the adverse influence of missingness on predictions and imputation of areal HIV new diagnoses. One is the likelihood-based method that integrates the missing mechanism into the likelihood function, and the other is a nonparametric algorithm for matrix factorization imputation. Numerical studies and the Philadelphia data analysis demonstrate that the two proposed methods can significantly improve prediction and imputation based on left-censored HIV data. We also compare the two methods on their robustness to model misspecification and find that both methods appear to be robust for prediction, while their performance for imputation depends on model specification.

KEYWORDS
left-censored, likelihood, matrix factorization, missing value, spatiotemporal data

1 | INTRODUCTION

As the world focuses on Covid-19, we should not forget other pandemics that plague our lives. Since the first known human case in 1959, HIV has infected more than 77 million people worldwide. Despite significant treatment progress, HIV has killed over 35 million people, including 690,000, last year alone. In the United States, the number of newly diagnosed cases of HIV has declined by 19% in the previous decade after years of continuous work to curb HIV. However, the progress has been uneven among geographical regions and demographic groups. The US Department of Health and Human Services (HHS) proposed a plan for the United States to end the HIV epidemic within 10 years. This initiative leverages critical scientific advances in HIV prevention, diagnosis, treatment, and outbreak response by coordinating the highly successful programs, resources, and infrastructure of many HHS agencies and offices. Proven interventions, including pre-exposure prophylaxis (PrEP), can effectively prevent new HIV transmissions.

The HIV new diagnosis prediction can help orchestrate the limited public health resources to priority areas and is vital for HIV intervention. Thanks to the increased accessibility to data at a finer scale, recent HIV predictions have been made at a higher resolution, such as at the county level or even zip code level (Chan et al., 2018; Shand et al., 2018; Sass et al., 2021). The higher resolution prediction provides a more informative guide for planning ahead to the local health care officials. However, HIV prediction is often plagued by missing data suppressed for privacy protection. For example, in the publicly available HIV new diagnosis data (https://aidsvu.org/local-data/#/), any rate, calculated as the number of HIV new diagnoses every 100,000 people, less than five is suppressed for confidentiality issues, which imposes challenges for fine spatial resolution

prediction. In the previous analysis, the suppressed data have often been ignored (Shand et al., 2018) or a single plausible number (Sass et al., 2021) has been imputed. Although Sass et al. (2021) studied the sensitivity of prediction to various imputed constants and found that different imputed constants induced variation of the prediction results, no further or rigorous investigation was conducted on how to integrate suppressed data into the analysis.

Our motivating data set is the zip-code level HIV new diagnosis counts in each quarter between 2009 to 2015. For privacy considerations, the Centers for Disease Control and Prevention (CDC) routinely suppress the data from regions with less than five new HIV diagnoses or less than 100 inhabitants for confidentiality. Since Philadelphia is heavily populated, all suppression is due to the first rule, making the observed HIV diagnoses left censored. To reduce the suppression rate, the zip code level data in Philadelphia were aggregated into 19 areal data by merging adjacent zip codes; see Figure 5. All counts of areal diagnoses less than or equal to 5 are suppressed. Ignoring these suppressed data may lead to bias in the statistical inference. For instance, dropping these small values may spuriously enlarge the mean estimators and shrink the variance estimator. Using the terminology in Zhao and Shao (2015), Kim and Yu (2011), Yuan and Yin (2010), Ibrahim et al. (2001), and Rubin (1976), if the missing mechanism is independent of the missing values, it is called ignorable or missing at random (MAR). Otherwise, it is called nonignorable or missing not at random (MNAR). MAR, if applicable, is widely assumed for simplifying the data structure (Camoni et al., 2013; Hall et al., 2008; Karon et al., 2008; Ndawinz et al., 2011; Pakianathan et al., 2018). However, many data carry nonignorable missing values in practice. Under MNAR, naively dropping the unobserved data or imputing missing values under MAR may bring bias and hurt the consistency of the estimations (Leacy et al., 2017).

We demonstrate the risk of dropping MNAR values using a simple example. We randomly sample $X_i, i = 1, \ldots, 100$ from $N(0,1)$ and then obtain $Y_i = 1 + X_i + \epsilon_i$ where $\epsilon_i \sim N(0,1)$ is the white noise. If we use all 100 pairs to fit a linear regression $Y = \alpha + \beta X$, the fitted model is unbiased to the true model as shown in Figure 1a. However, if we suppress the 26 pairs whose Y value is less than 0, and only use the rest data to fit the linear model, the fitting results in Figure 1b show obvious bias in both the intercept α and slope β. Biased estimation often implies biased prediction.

In general, existing methods for handling informative missing values in epidemic data (e.g., Bärnighausen et al., 2011; Marra et al., 2017) fall into two categories, likelihood-based inference and imputation. The likelihood-based methods focus on simultaneously modeling the observed response values and missing mechanisms. More specifically, let $\boldsymbol{y}$ be the observed responses, $\boldsymbol{r}$ the missing indicator, and $\boldsymbol{x}$ the covariates. Rather than modeling $(\boldsymbol{y}|\boldsymbol{x},\theta)$, where θ is a set of unknown parameters, the likelihood-based approaches focus on modeling $(\boldsymbol{r},\boldsymbol{y}|\boldsymbol{x},\theta)$. The inclusion of missing mechanisms in the likelihood can help reduce the bias caused by missingness. The imputation methods handle the missingness in a different way. First, the missing values are imputed, and then the data with imputed values are treated as the full data. The imputation can be performed in either a single or multiple fashion (e.g., Murray & Reiter, 2016; Si & Reiter, 2013), but all are mainly based on local observations such as the mean or mode of the "neighbors" regardless of the imputation type. For the Philadelphia data, since all observed values are larger than 5 while the missing values are less than or equal to 5, imputation using observed neighbors alone will likely produce values larger than 5. Thus, directly applying the traditional imputation to the suppressed HIV diagnoses may carry bias. Imputing such left-censored data has been an interesting topic for public health data analysis (e.g., Canales et al., 2018; Erdman et al., 2021; Quick, 2019; Wei et al., 2018) and environmental data analysis (e.g., Cohen Jr, 1950; Gleit, 1985; Kucharska et al., 2022; Mohamed et al., 2021; Sahoo & Hazra, 2021).

Specifically for censored spatiotemporal data, as our Philadelphia new diagnoses, many statistical methods have been developed to take spatial dependency into account in either likelihood-based inferences or imputation; see Canales et al. (2018), Wei et al. (2018), Erdman et al. (2021), Kucharska et al. (2022), Mohamed et al. (2021), Quick (2019), Sahoo and Hazra (2021), and Zhou and Hanson (2018). Some machine learning methods can also be used to handle censored data; see Spooner et al. (2020), Vock et al. (2016), and Wang and Zhou (2017). However, these machine learning methods are designed mainly for independent time-to-event data rather than spatiotemporal data.

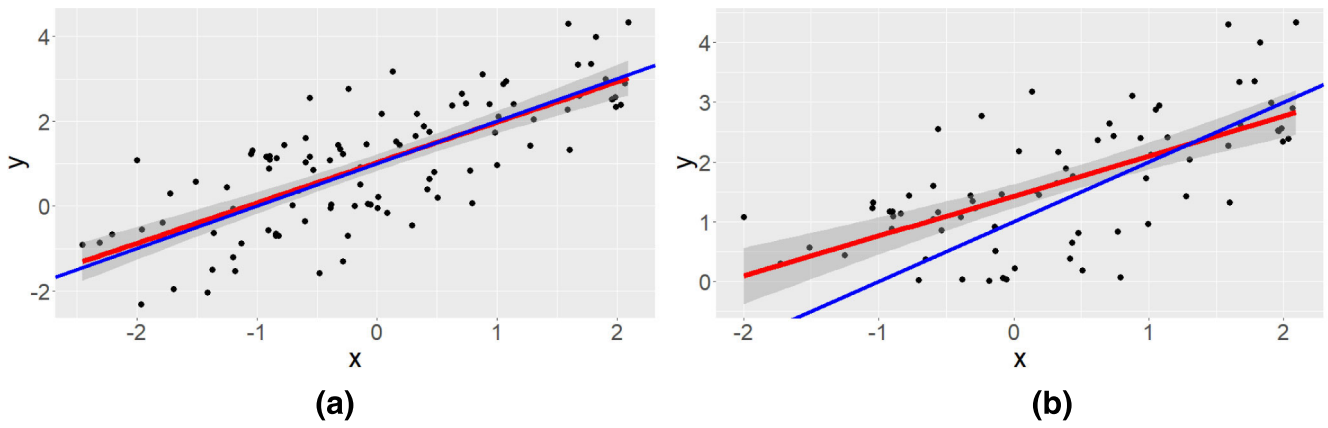


FIGURE 1 Linear regression with (a) the entire data, or (b) only the data having positive Y. The red line is the fitted model and the blue line is the true model.

We propose two approaches to reduce the estimation and prediction bias caused by missing HIV diagnoses. One is a traditional likelihood-based method, and the other adapts a machine learning method, matrix factorization (MF) imputation, to our data. We also derive an expectation-maximization (EM) algorithm for the likelihood method to accelerate its computation.

The paper is organized as follows. In Section 2, we focus on the likelihood method and its EM-algorithm development, while in Section 3, we introduce an adapted MF method for imputation. Section 4 presents simulation results to evaluate the performance of the two approaches relative to other methods. In Section 5, we apply the two proposed methods with two competing methods to the Philadelphia new diagnosis data. Section 6 concludes the paper with a brief discussion.

2 | LIKELIHOOD-BASED APPROACH

Likelihood methods require the specification of the joint distribution of response variables and missing indicators. Either selection models or pattern-mixture models can usually accomplish this specification; see Little (1993) and Rubin (1976). The selection method first models the complete response values, including both the observed and the unobserved values, and then models the missing indicators condition on the complete response values (Diggle & Kenward, 1994). In contrast, the pattern-mixture method models the missing mechanism first and then models the conditional distribution of the response values given the missingness (Little & Yau, 1996). For the HIV new diagnosis data, the conditional distribution of missing indicators given the complete response values degenerates into a deterministic function due to the data suppressing rule; that is, only the values larger than five will be observed and otherwise suppressed. This makes the selection approach convenient for this data type. Below we detail our joint model by following the two steps of the selection method.

2.1 | Spatiotemporal model for complete data

We first specify the model for the complete data. Disease incidence accounts are in general modeled by a Poisson or binomial distribution whose mean depends on the standardized relative risk, which can further be modeled by linear mixed models. Altogether, the modeling of disease accounts is often put in the framework of generalized linear mixed models (Breslow & Clayton, 1993). However, if the counts are relatively large, a transformed Gaussian model can be a more flexible choice (Arnold et al., 1999) because it avoids the potential overdispersion/underdispersion issue inherent in the Poisson or binomial distribution. The transformed Gaussian also makes it easy to model the dependence structure often exhibited in spatiotemporal public health data. Gaussian distribution has been justified for modeling the HIV data after the log transformation (Shand et al., 2018). We examined the Gaussianity of the log-transformed Philadelphia data in Figure 2. The plot also shows that the logarithm of our HIV new diagnosis data in Philadelphia approximately follows a normal distribution. We therefore model our data using a log Gaussian process.

Let $y_{i,j}$ denote the new diagnosis at time i and location j, and let $z_{i,j} = \log(y_{i,j})$. We focus on modeling $z_{i,j}$. Let $\boldsymbol{Z} = [z_{i,j}]_{i=1,\ldots,I;j=1,\ldots,J}$ be the vector of the logarithmic HIV new diagnoses at all locations and all time points during the period of interest. Let $\boldsymbol{x}_{i,j}$ denote the covariates for modeling

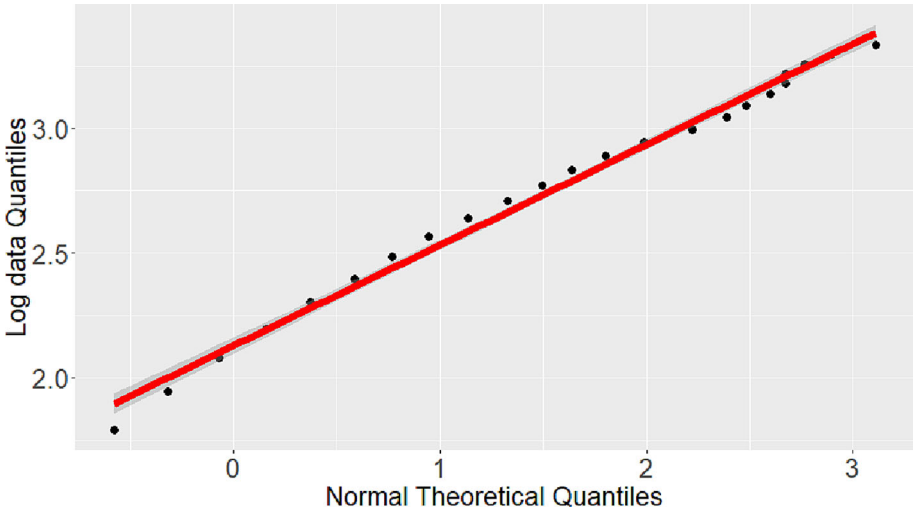


FIGURE 2 Q-Q plot between the standard normal distribution and the empirical distribution of the log transformed HIV new diagnoses in Philadelphia.

$z_{i,j}$, and let $\boldsymbol{X} = [\boldsymbol{x}_{i,j}]_{i=1,\ldots,I;j=1,\ldots,J}$ be the matrix of covariates at all locations and time points. We further divide $\boldsymbol{Z}$ into the missing component, denoted by $\boldsymbol{Z}_M$, and the observed component, denoted by $\boldsymbol{Z}_O$.

We model the space-time new diagnosis values, $\boldsymbol{Z}$, as

$$z_{i,j} = \eta(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}) + \epsilon_{i,j},$$

where η is a deterministic function with parameters $\boldsymbol{\theta}$ and $\epsilon = [\epsilon_{i,j}]_{i=1,\ldots,I;j=1,\ldots,J} \sim N(0,\Sigma)$ for a covariance matrix Σ. Then we have the distribution for $\boldsymbol{Z}$ as

$$\text{pr}(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta},\Sigma) = \phi(\boldsymbol{Z};\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{X}),\Sigma), \tag{2.1}$$

where $\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{X}) = [\eta(\boldsymbol{\theta},x_{i,j})]_{i=1,\ldots,I;j=1,\ldots,J}$ is the mean vector of $\boldsymbol{Z}$ with length $I \times J$, and $\phi(\cdot;\boldsymbol{\mu},\Gamma)$ is the density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Γ.

The covariance matrix Σ can be governed by any valid space-time covariance function (e.g., Choi et al., 2013; Gneiting, 2002; Shand & Li, 2017). For simplicity, we assume the covariance matrix follows a space-time separable model, which is perhaps the most popular choice for spatiotemporal data applications. More specifically,

$$\Sigma(\sigma,\rho,\lambda) = \sigma^2 \Sigma_s(\rho) \bigotimes \Sigma_t(\lambda), \tag{2.2}$$

where σ^2 is the variance parameter, $\Sigma_s(\rho)$ is a J by J matrix describing the spatial correlation, and $\Sigma_t(\lambda)$ is an I by I matrix for temporal correlation. To investigate which correlation structure is appropriate for $\Sigma_s(\rho)$, we consider three distinctive structures commonly used for environmental data. One is the exchangeable model for which all diagonal elements of $\Sigma_s(\rho)$ are one while off-diagonal elements are ρ. This model assumes a homogeneous correlation between any two different locations. The second is the exponential spatial correlation function, i.e., $\Sigma_{s,i,j}(\rho) = \exp(-\rho d_{i,j})$, where $d_{i,j}$ is the Euclidean distance between the ith and jth locations. This model assumes the correlation decays at a rate ρ as the distance increases and is widely used for geospatial data. The last is the conditional autoregressive (CAR) model (Leroux et al., 2000; Shand et al., 2018), for which $\Sigma(\rho) = \{(1-\rho)\mathbf{I} + \rho\mathbf{R}\}^{-1}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{R}$ denotes a neighborhood matrix with the i th diagonal element as the total number of neighbors for location i and the (i,j)th off-diagonal element as -1 if locations i and j share a border and 0 otherwise. The CAR model assumes that the correlation among observations arises from their being neighbors and is generally considered the most appropriate model for spatially aggregated data such as county or zip code level data. We assume $\Sigma_t(\lambda)$ is governed by an autoregressive model of order 1 with the autoregressive parameter λ.

Typically, the mean function $\eta(\boldsymbol{\theta},\boldsymbol{x}_{i,j})$ is assumed to be a linear function of covariates. Thus, space-time varying covariates can naturally describe how the mean varies over locations and time points. However, the demographic data that can be used as covariate for our HIV diagnoses are collected only every 10 years; thus, a linear function of $x_{i,j}$ can only help capture the spatial variability but not the desired temporal variability, that is, $x_{i_1,j} = x_{i_2,j}$. Under this circumstance, we treat the means at every location as unknown parameters $\theta_j = \eta(\boldsymbol{\theta},\boldsymbol{x}_{i,j})$, which is more general than assuming a linear function of covariates. We assume the mean is invariant over time, as there is no obvious trend of the observations. However, η can be easily replaced by a function of covariates whenever helpful. Thus, the following of this section continues the derivation based on the general case of $\eta(\boldsymbol{\theta},\boldsymbol{X})$.

2.2 | Joint model with missing indicators

For each time i and location j, let $r_{i,j}$ be the missing indicator defined on (i,j) such that $r_{i,j} = 1$ when the new diagnosis $y_{i,j}$ and thus $z_{i,j}$ are missing, and $r_{i,j} = 0$ otherwise. In our Philadelphia zip-code level data, the missing indicator is absolutely determined by whether the positive HIV count is greater than 5. Let $\boldsymbol{A}$ denote the set where $z_{i,j}$ is suppressed. Then the missing indicator can be written as $r_{i,j} = \mathbf{1}(z_{i,j} \in \boldsymbol{A})$, where $\mathbf{1}(\cdot)$ is the indicator function. Let $\boldsymbol{R} = [r_{i,j}]_{i=1,\ldots,I;j=1,\ldots,J}$ be the vector of missing indicators for all locations and time points. The conditional distribution of $\boldsymbol{R}$ given complete response values and covariates is

$$\text{pr}(\boldsymbol{R}|\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\theta},\Sigma) = \text{pr}(\boldsymbol{R}|\boldsymbol{Z}) = \prod_{i=1}^{I}\prod_{j=1}^{J} \mathbf{1}(z_{i,j} \in \boldsymbol{A})^{r_{ij}} \mathbf{1}(z_{i,j} \in \boldsymbol{A}^c)^{1-r_{ij}}, \tag{2.3}$$

where $\boldsymbol{A}^c$ is the complement of set $\boldsymbol{A}$. Since the missing mechanism for the HIV data is known, the conditional probability in (2.3) degenerates to either 1 or 0.

Then the likelihood of unknown parameters based on both the observed responses and missing indicators is

$$\begin{aligned} &L(\boldsymbol{\theta},\Sigma|\boldsymbol{R},\boldsymbol{Z}_O) \\ &= \phi(\boldsymbol{Z}_O;\boldsymbol{\eta}_O(\boldsymbol{\theta},\boldsymbol{X}_O),\Sigma_O)\int_{\boldsymbol{Z}_M \in \boldsymbol{A}} \text{pr}(\boldsymbol{Z}_M|\boldsymbol{Z}_O,\boldsymbol{X},\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M, \end{aligned} \tag{2.4}$$

where the subscript O refers to the component corresponding to the observations and M to the missing values. The derivation of (2.4) is deferred to Appendix A.1. Since $\text{pr}(\boldsymbol{Z}_M,\boldsymbol{Z}_O|\boldsymbol{X},\boldsymbol{\theta},\Sigma)$ is assumed to follow a joint normal distribution, we can easily derive $\text{pr}(\boldsymbol{Z}_M|\boldsymbol{Z}_O,\boldsymbol{X},\boldsymbol{\theta},\Sigma)$, which also follows a joint normal distribution with means and covariance matrix that depend only on $\boldsymbol{Z}_O,\boldsymbol{X},\boldsymbol{\theta}$, and Σ. The purpose of Equation (2.4) is to express the likelihood function into a function of known distributions so that the likelihood can be calculated for parameter estimation. Essentially, the estimation of Σ is equivalent to estimating σ^2,ρ, and λ.

2.3 | Monte Carlo EM algorithm

The integration in the likelihood function (2.4) makes it hard to derive closed-form maximum likelihood estimators (MLE) for the unknown parameters. The expectation-maximization (EM) algorithm (Dempster et al., 1977), an iterative method to find MLE when the model depends on missing or latent variables, has been widely used. The popularity of the EM algorithm is gained by its easy implementation and numerical stability. Furthermore, the EM algorithm can converge under weak assumptions and the convergence is usually fast. We develop an EM algorithm to find MLEs for our intricate likelihood function.

The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated at the current estimate of the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E-step. In the E-step, define $Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right)$ as the conditional expectation of the log-likelihood over the missing values, where $\boldsymbol{\theta}^{(t)}$ and $\Sigma^{(t)}$ are the estimates of $\boldsymbol{\theta}$ and Σ at step t. Particularly for likelihood function (2.4), the Q-function is defined as

$$\begin{aligned} Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right) &= \text{E}_{\boldsymbol{Z}_M|\boldsymbol{Z}_O,\boldsymbol{R},\boldsymbol{\theta}^{(t)},\Sigma^{(t)}}[\log\{L(\boldsymbol{\theta},\Sigma|\boldsymbol{Z}_M,\boldsymbol{Z}_O)\}] \\ &= \int \text{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta},\Sigma)\log\{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\}d\boldsymbol{Z}_M. \end{aligned} \tag{2.5}$$

In the M-step, we update $(\boldsymbol{\theta},\Sigma)$ by $(\boldsymbol{\theta}^{(t+1)},\Sigma^{(t+1)}) = \text{argmax}Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right)$. We repeat these two steps until the estimates $\boldsymbol{\theta}^{(t)}$ and $\Sigma^{(t)}$ converge.

The calculation of the Q function is not trivial. We first derive that the conditional distribution of the missing values given the observed values and missing indicators, $\text{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta},\Sigma)$, follows a multivariate truncated normal distribution:

$$\text{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta},\Sigma) = \frac{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\prod_{(i,j)\in\mathcal{M}}\mathbf{1}\left(Z_{ij}\in(-\infty,\log(5)]\right)}{\text{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma)}, \tag{2.6}$$

where $\mathcal{M}$ is the index sets of the missing data. The derivation of (2.6) is deferred to Appendix A.1.1. Then we obtain

$$\begin{aligned} &Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right) \\ &= \frac{\int_{\boldsymbol{Z}_M\in(-\infty,\log(5)]}\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\log\{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\}d\boldsymbol{Z}_M}{\int_{\boldsymbol{Z}_M\in(-\infty,\log(5)]}\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M}. \end{aligned} \tag{2.7}$$

Again, the derivation of (2.7) is deferred to Appendix A.1.2. We still have integration in both the numerator and denominator of (2.7) that leads to a no closed form solution. To bypass this difficulty in the E-step, we employ the Monte Carlo EM (MCEM) algorithm proposed by Wei and Tanner (1990).

With MCEM, rather than calculating the complicated analytic form of the expectation, we generate m independent samples, $\boldsymbol{Z}_{M,1},\ldots,\boldsymbol{Z}_{M,m}$, based on $(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta}^{(t)},\Sigma^{(t)})$. Then the Q function can be approximated by the sample mean:

$$\begin{aligned} Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right) &\approx \frac{1}{m}\sum_{i=1}^{m}\log\{L(\boldsymbol{\theta},\Sigma|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{Z}_{M,i})\} \\ &= \frac{1}{m}\sum_{i=1}^{m}\log\{\phi(\boldsymbol{Z}_{M,i},\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\}. \end{aligned} \tag{2.8}$$

Since $\boldsymbol{Z}_{M,i}, i = 1,\ldots,m$ are truncated Gaussian processes as shown in (2.6), we adopt a direct and efficient Gibbs sampler introduced by Li and Ghosh (2015) to sample a truncated Gaussian process through univariate truncated normal random variables. In the ith iteration of the Gibbs sampler, we sample the kth element of $\boldsymbol{Z}_M, Z_M^k$, from the conditional distribution $(Z_M^k|\boldsymbol{Z}_M^{-k},\boldsymbol{Z}_O,\boldsymbol{R},\boldsymbol{\theta},\Sigma)$, where $\boldsymbol{Z}_M^{-k}$ is the rest of $\boldsymbol{Z}_M$ except Z_M^k. This conditional distribution is simply a truncated univariate normal distribution:

$$\mathrm{pr}(Z_M^k|\boldsymbol{Z}_M^{-k},\boldsymbol{Z}_O,\boldsymbol{R},\boldsymbol{\theta},\Sigma) \propto \mathbf{1}\left(Z_M^k \in (-\infty,\log 5]\right)\mathrm{pr}(Z_M^k|\boldsymbol{Z}_M^{-k},\boldsymbol{Z}_O,\boldsymbol{\theta},\Sigma).$$

When incorporating a Gibbs sampler in the MCEM algorithm, once the MCEM algorithm is about to converge, the difference between $(\boldsymbol{\theta}^{(t)},\Sigma^{(t)})$ and $(\boldsymbol{\theta}^{(t+1)},\Sigma^{(t+1)})$ is small; that is, the difference between $(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta}^{(t)},\Sigma^{(t)})$ and $(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta}^{(t+1)},\Sigma^{(t+1)})$ is small. We therefore use the sampled value of the previous iteration as the initial value for the Gibbs sampler of the next iteration, which leads to very fast convergence.

In the M-step, to avoid a high-dimensional optimization problem, we update $\boldsymbol{\theta}$ and Σ in the Q function in (2.5) sequentially. First, we update Σ by $\mathrm{argmax}_\Sigma Q\left(\boldsymbol{\theta}^{(t)},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right)$. Then, given $\Sigma^{(t+1)}$, we update $\boldsymbol{\theta}$ by $\mathrm{argmax}_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta},\Sigma^{(t+1)}|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right)$, which essentially reduces to a weighted least square problem. Specifically, by (2.8), Q in the second step can be written as

$$Q\left(\boldsymbol{\theta},\Sigma^{(t+1)}|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right) = \frac{1}{m}\sum_{i=1}^{m}\log\{\phi(\boldsymbol{Z}_{M,i},\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma^{(t+1)})\}.$$

Given $\Sigma^{(t+1)}$, maximizing Q is equivalent to maximizing the likelihood of m i.i.d. multivariate normal distribution samples with unknown parameters $\boldsymbol{\theta}$.

2.4 | Prediction

Having estimated the model parameters by the likelihood-based method, we then use kriging to predict future HIV diagnoses. Kriging, introduced in Cressie (2015), is the most fundamental prediction method for spatial and spatiotemporal data. Essentially, kriging prediction is the conditional mean at the current space-time location given the observations at other locations. Let $\boldsymbol{Y}_P$ be the vector of HIV diagnoses that need to be predicted and $\boldsymbol{Z}_p = \log(\boldsymbol{Y}_P)$, where the logarithm is taken element-wise. If there are no missing data, we can easily derive the conditional multivariate normal distribution of $\boldsymbol{Z}_P$ and thus the conditional distribution of $\boldsymbol{Y}_P$ given $(\boldsymbol{Z}_O,\boldsymbol{Z}_M,\hat{\Sigma},\hat{\boldsymbol{\theta}})$. Then we use the mean of the conditional distribution of $\boldsymbol{Y}_P$ as the prediction.

However, since $\boldsymbol{Z}_M$ is missing, and instead we only observe the missing indicator $\boldsymbol{R}$, the conditional distribution of $\boldsymbol{Z}_P$ becomes

$$\mathrm{pr}(\boldsymbol{Z}_P|\boldsymbol{R},\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}}) = \int \mathrm{pr}(\boldsymbol{Z}_P,\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}})d\boldsymbol{Z}_M, \tag{2.9}$$

for which the integration is difficult to calculate. We again resort to the Monte Carlo method to bypass the integration. Instead of generating $\boldsymbol{Z}_P$, we generate $(\boldsymbol{Z}_P,\boldsymbol{Z}_M)$ simultaneously. We first write the integrand of (2.9) into the following:

$$\mathrm{pr}(\boldsymbol{Z}_P,\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}}) = \mathrm{pr}(\boldsymbol{Z}_P|\boldsymbol{Z}_M,\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}})\mathrm{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}}).$$

Then we draw samples in two steps. The first step is to sample $\boldsymbol{Z}_{M,1},\ldots,\boldsymbol{Z}_{M,n}$ from $\mathrm{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}})$ using the MCEM algorithm, as discussed in Section 2.3. The second step is to sample $\boldsymbol{Z}_{P,i}$ from $\mathrm{pr}(\boldsymbol{Z}_P|\boldsymbol{Z}_{M,i},\boldsymbol{Z}_O,\hat{\Sigma},\hat{\boldsymbol{\theta}})$, a conditional multivariate normal distribution for $i = 1,\ldots,n$. Then the sample mean of logarithm of $\boldsymbol{Z}_P, \overline{\boldsymbol{Y}}_P = \frac{1}{n}\sum_{i=1}^{n}\exp(\boldsymbol{Z}_{P,i})$, is taken as the prediction of $\boldsymbol{Y}_P$.

It is worth noting that once the MCEM sampling algorithm converges, we can treat the sampled $\boldsymbol{Z}_{M,1},\ldots,\boldsymbol{Z}_{M,n}$ as the possible realizations of the missing values. The sample mean of the logarithm of $\boldsymbol{Z}_M, \overline{\boldsymbol{Y}}_M = \frac{1}{n}\sum_{i=1}^{n}\exp(\boldsymbol{Z}_{M,i})$ can be considered the imputation of missing values $\boldsymbol{Y}_M$. Then the prediction can be made based on the complete data, which comprises both the observed and imputed values. Such imputation serves as an alternative approach to making predictions by taking advantage of the byproduct of likelihood estimates.

3 | MF IMPUTATION

This section considers imputation as an alternative strategy for handling missing values in the HIV data. MF is a well-known machine learning method for missing value imputation. It can take advantage of the intrinsic correlation structure of the dependent data to provide reliable imputations. When applied to spatiotemporal data, the data can be formed as a matrix according to the spatial dimension and the temporal dimension.

However, although MF imputation works well for imputing MAR spatiotemporal data in Huang et al. (2013), Ranjbar et al. (2015), and Yang et al. (2021), it is not well suited for the censored data (e.g., Freeman et al., 2022; Moosavi et al., 2021; Rebouillat et al., 2021). In our data, all observations are at least six while the missing values are smaller or equal to five, directly applying the traditional MF will yield inflated imputations as with other imputation methods. To enforce the imputed values being within [0,5] for our data or $[a,b]$ for a more general case, we propose a penalized MF to accommodate the constraints of missing values based on Takács et al. (2008). This idea comes from constrained optimization, in which the original constrained optimization question can be converted to an unconstrained optimization question with a penalty term; see Yeniay (2005).

Let $\boldsymbol{D}$ be the data matrix of logarithmic HIV new diagnoses, including the missing values. Since our data are spatiotemporal, we naturally formulate the data into a matrix with rows representing time and columns representing the spatial locations. The MF method in Takács et al. (2008) seeks a $\hat{\boldsymbol{D}}$ such that $\boldsymbol{D} \approx \hat{\boldsymbol{D}} = \boldsymbol{P}^T\boldsymbol{Q}$, that is, $\hat{d}_{i,j} = \boldsymbol{p}_i^T\boldsymbol{q}_j$, where $\hat{d}_{i,j}$ is the (i,j)th element of $\hat{\boldsymbol{D}}$, $\boldsymbol{p}_i$ the ith column of $\boldsymbol{P}$, and $\boldsymbol{q}_j$ the jth column of $\boldsymbol{Q}$. Then a straightforward method to derive $\hat{\boldsymbol{D}}$ is to find the $\boldsymbol{P}$ and $\boldsymbol{Q}$ that minimize $\|\boldsymbol{D} - \boldsymbol{P}^T\boldsymbol{Q}\|_F$, where $\|\boldsymbol{X}\|_F$ is the Frobenius norm of a matrix $\boldsymbol{X}$. The basic $\hat{\boldsymbol{D}}$ form can be generalized by adding bias terms such as $\hat{d}_{i,j} = \boldsymbol{p}_i^T\boldsymbol{q}_j + e_i + f_j + c$, where e_i, f_j, and c represent time and location specific bias parameters and an overall shift, respectively. To ensure the identifiability of each bias parameter, c is typically fixed at the mean value of the whole data matrix. In practice, penalties on the norm of each element in $\hat{\boldsymbol{D}}$, $\|\boldsymbol{P}\|_F + \|\boldsymbol{Q}\|_F + \|\boldsymbol{e}\|_2 + \|\boldsymbol{f}\|_2$, will be imposed to the loss function, where $\boldsymbol{e} = [e_i]_{i=1,\ldots,I}$, $\boldsymbol{f} = [f_j]_{j=1,\ldots,J}$, and $\|\boldsymbol{x}\|_2$ is the L^2 norm of a vector $\boldsymbol{x}$. This penalty can further help improve the identifiability of the optimization procedure and avoid $\boldsymbol{P}$ and $\boldsymbol{Q}$ being coherent matrices.

If the missing values are expected to reside in a region $[a,b]$, particular $(-\infty, \log(5)]$ for our suppressed data, we propose to impose another penalty $(a - \hat{d}_{i,j})_+^2 + (\hat{d}_{i,j} - b)_+^2$, where $(x)_+ = \max\{x, 0\}$, for all missing $d_{i,j}$ to enforce the imputed missing values being within the desired range. The final loss function given a and b becomes

$$\begin{aligned} L(\boldsymbol{P},\boldsymbol{Q}) = & \sum_{(i,j)\in\mathcal{O}} \left(d_{i,j} - \boldsymbol{p}_i^T\boldsymbol{q}_j - e_i - f_j - c\right)^2 \\ & + \eta \sum_{(i,j)\in\mathcal{M}} \left\{ (a - \boldsymbol{p}_i^T\boldsymbol{q}_j - e_i - f_j - c)_+^2 + (\boldsymbol{p}_i^T\boldsymbol{q}_j + e_i + f_j + c - b)_+^2 \right\} \\ & + \zeta(\|\boldsymbol{P}\|_F + \|\boldsymbol{Q}\|_F + \|\boldsymbol{e}\|_2 + \|\boldsymbol{f}\|_2), \end{aligned} \tag{3.1}$$

where $\mathcal{O} = \{(i,j) : d_{i,j} \text{ is observed}\}$ and $\mathcal{M} = \{(i,j) : d_{i,j} \text{ is missing}\}$. Note that the penalty is well defined with $a = -\infty$ or $b = \infty$. In those two cases, $a - \boldsymbol{p}_i^T\boldsymbol{q}_j - e_i - f_j - c$ or $\boldsymbol{p}_i^T\boldsymbol{q}_j + e_i + f_j + c - b$ is negative, and then takes 0 after going through the $(x)_+$ operator.

To search the optimizer of this loss function, we use stochastic gradient descent (SGD) with a specified learning rate. Compared to other optimization methods, SGD is more computationally efficient and easy to implement. As shown in the Appendix B1, each iteration in the SGD has simple form gradients that are convenient to calculate. There are two tuning parameters, η and ζ, in our loss function. We simply set η a large value to enforce the imputed values to be within the range of $[a,b]$. There is no theoretical or intuitive way to choose the tuning parameter ζ, and cross-validation is not feasible here as we only have one sample. We use a grid search to select the ζ that minimizes the 1-step forecasting mean squared error.

We take the elements in $\hat{\boldsymbol{D}}$ corresponding to missing values as our imputation. Then we treat the data with imputed values as the complete data, apply the model (2.1) to estimate the parameters using the maximum likelihood method, and make predictions using kriging.

4 | SIMULATION STUDY

We conduct simulation studies to evaluate the prediction performance of the likelihood-based method and the MF imputation. Since the likelihood method requires a model assumption, its performance may be sensitive to model misspecification. We therefore compare these two methods under both scenarios of having a correctly specified model and a misspecified model in the likelihood function.

4.1 | Simulation settings

In the correctly specified case, we first generate $\tilde{\boldsymbol{Z}}$ following model (2.1) with $\eta(\boldsymbol{\theta}, x_{i,j}) = \theta_j, j = 1,\ldots,J$. Then we take $\lfloor \exp(\tilde{\boldsymbol{Z}}) \rfloor$ as the simulated new diagnosis counts, where $\lfloor a \rfloor$ (floored a) is the largest integer that is less than or equal to a. We carefully select coefficients $(\theta_1,\ldots,\theta_J)$ such that the simulated data have a similar mean value as the real data. For the covariance matrix Σ in (2.1), we try all three correlation structures mentioned in Section 2.1, which are exchangeable, exponential, and CAR model to generate data, respectively. For each correlation structure, we use two sets of covariance parameters (σ,ρ,λ) defined in (2.2). One represents a stronger spatiotemporal correlation, while the other represents a weaker correlation. Details of parameter settings for $(\theta_1,\ldots,\theta_J)$ and (σ,ρ,λ) are shown in Tables C1 and C2 in Appendix C1. For these simulated data, we use model (2.1) with the same correlation structure as in the data generation to estimate parameters.

In the misspecified case, we use Poisson marginal with Gaussian copula to generate data. Specifically, we first generate $\boldsymbol{U} \sim \mathbf{N}(\mathbf{0}, \tilde{\Sigma})$, where $\tilde{\Sigma}$ is the correlation matrix corresponding to the covariance matrix, Σ, in (2.2). Similar to the correctly specified case, we employ the three different correlation structures and two parameter settings in the correlation matrix $\tilde{\Sigma}$. Let $P_{\tau_{ij}}$ be the Poisson cumulative distribution function with parameter τ_{ij}. To make the Poisson random variables have a similar expectation as the log-normal distribution with parameters (θ_{ij}, σ^2) in the correctly specified case, we set $\tau_{ij} = \exp\left(\theta_{ij} + \frac{\sigma^2}{2}\right)$. Then we generate new HIV diagnosis counts $\tilde{Y}_{ij} = P^{-1}_{\tau_{ij}}(\Phi(u_{ij}))$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and u_{ij} is the element of $\boldsymbol{U}$ at time i and location j. For these simulated data, we again use model (2.1) to estimate parameters, which is different from the model used for data generation. Furthermore, we intentionally choose different correlation functions for estimation than for data generation to exacerbate model misspecification. For example, if the exchangeable correlation structure is used for data generation, then we choose an exponential correlation or CAR model for estimation.

We generate data at 19 locations and 28 time points, following the size of the Philadelphia HIV new diagnosis counts. Any values less than or equal to 5 in the simulated data are suppressed. To study the performance of the methods with a larger I, the number of time points, we additionally simulate data with $I = 61$ for the scenarios where the correlation parameters represent a stronger spatiotemporal correlation. In summary, for both the correctly and the misspecified cases, we generate data under nine scenarios formed by pairing each of the three correlation structures with each of the three settings: Strong correlation with small sample size, weak correlation with small sample size, and strong correlations with large sample size.

To compare prediction performance, we use all data but those at the last time point to fit the model and then make predictions for the last time point. The root mean squared error (RMSE) and the bias of the predictions are used as the assessment metrics. We handle missing values in model fitting based on five approaches, including our proposed likelihood-based method and MF imputation, two baseline methods of constant imputation (CtI) and ignore missing (IgM), and the oracle method for which the complete data with no missing values is hypothetically available. CtI imputes missing values using a constant 2.5, the median of the missing interval [0,5], while IgM simply ignores missing values and considers the observations as the complete data. All approaches but the likelihood-based method first fit model (2.1) using the classical likelihood method based on the data with or without imputations, and then make predictions using the traditional kriging method. All model fittings employ the same covariance model, whether correctly or misspecified. For each model and parameter setting, we run the simulation 200 times.

4.2 | Simulation results

We use the MCEM algorithm described in Section 2.4 to estimate parameters and make predictions for the likelihood-based method. In the E-step, we run 20,000 steps for the Gibbs sampler, drop the first 5000 as burn-in, and then take one every 200 for thinning. In our simulation, the E-M algorithm converges in several steps, and we thus set eight iterations to guarantee convergence. In implementing the MF method, we pick a relatively large value for $\eta, \eta = 100$, in the MF loss function (3.1). We also set $\zeta = 0.01$ for all simulations because a grid search of ζ for each single simulation is computationally extensive. The choice of 0.01 is based on a grid search for the Philadelphia data and we find the results are insensitive to small variations of ζ. For the SGD in the MF approach, we set the learning rate as 0.001 and run 3000 epochs.

We first examine the imputation RMSE in Figure 3. Both the likelihood and MF provide better imputations than the conventional CtI. If the model is correctly specified, the imputation byproduct from the likelihood-based method seems to carry smaller RMSE than the MF imputation, while under the misspecified model, the likelihood-based imputation has a larger RMSE than the MF imputation. The imputation bias deferred to Figure D1 in Appendix D shows a similar pattern as the imputation RMSE. The constant imputation method leads to the largest bias. The likelihood method provides a lower bias than the MF when the model is correctly specified, while the reverse is true when the model is misspecified.

The prediction RMSE of all models and at all parameter settings are reported in Figure 4. Surprisingly, the likelihood and MF predictions both achieve nearly the same RMSE as the oracle method with no missing values at all different combinations of generating and fitting models, regardless of the likelihood function being correctly or misspecified. This finding shows that prediction, compared to imputation, is less sensitive to model specification. CtI and Igm are apparently less competitive than the likelihood and MF in terms of prediction. Figure 4 also indicates that when the spatiotemporal correlation is weaker (parameter set 2), the RMSE becomes larger for all methods. This is not surprising as a weaker correlation indicates less information from neighbors can be borrowed when making kriging predictions. Comparing the prediction RMSE between $I = 28$ and $I = 61$ under both correctly specified and misspecified cases, we find the sample size 28 in our data is sufficient for attaining reliable prediction as a larger sample does not seem to reduce the prediction RMSE further. The comparison between different approaches purely reflects how different methods of handling missing values affect prediction because the modeling fitting and kriging operation are all based on the same spatiotemporal model.

We also examine the prediction bias in Figure D2 in Appendix D. Without surprise, the IgM generally has a large prediction bias. The constant imputation disturbs the original data dependency structure so the estimation and prediction based on this imputed data are spurious. That is why we observe large prediction RMSE in Figure 4. However, the prediction bias from this method can be small by chance, so the bias of CtI predictions appears volatile. The likelihood and MF methods consistently reduce the prediction bias compared to the IgM. In general, the likelihood-based method leads to a smaller prediction bias when the model is correctly specified, while MF leads to a smaller bias when the model is misspecified.

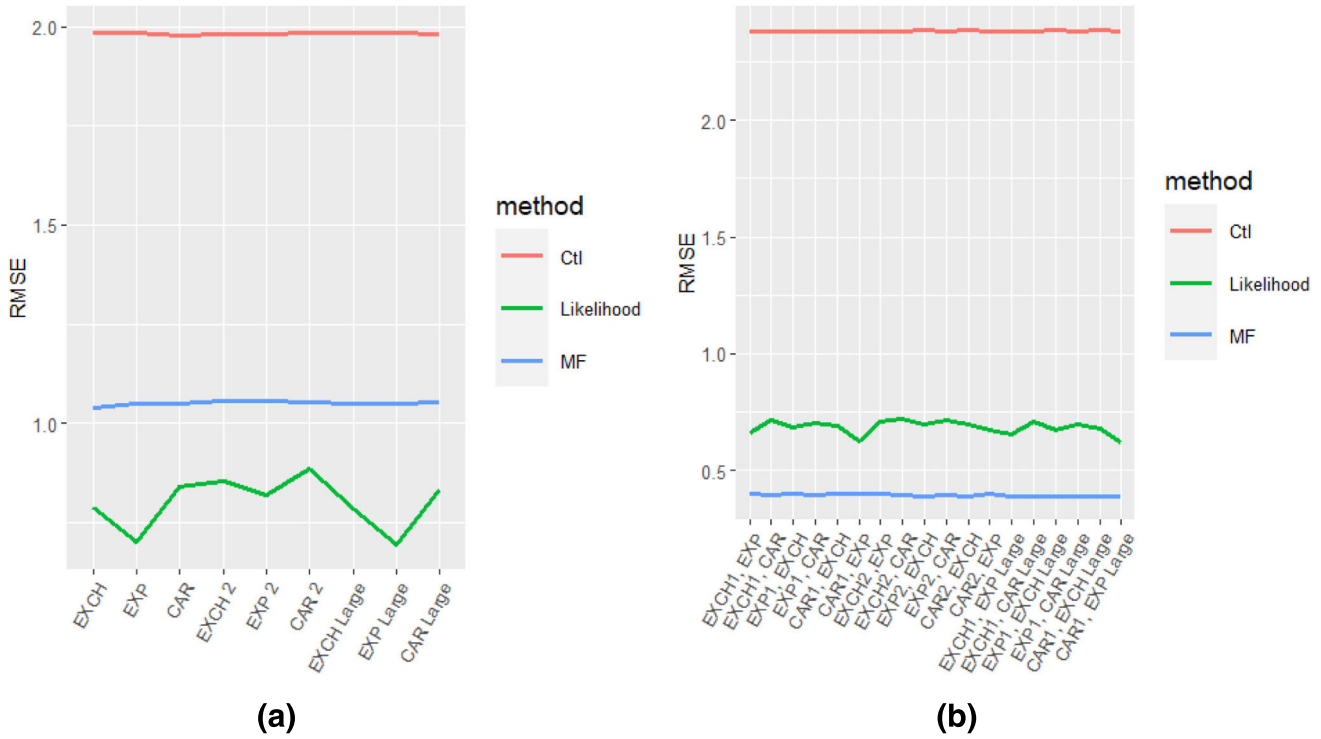


FIGURE 3 Imputation RMSE over all missing values from the constant imputation (Ctl), likelihood based method and MF imputation under different settings for misspecified model. The x-label has two model names. The first represents the fitting model, and the second is the data generating model. In both x-labels, "EXCH," "EXP," and "CAR" are exchangeable, exponential and CAR covariance models; "1" and "2" represent stronger and weaker spatiotemporal correlation; and "L" means $T = 61$ for the simulated data.

5 | APPLICATION TO PHILADELPHIA HIV NEW DIAGNOSES

The Philadelphia HIV new diagnosis data at zip code level contains quarterly counts of HIV new diagnoses from 2009 to 2015. All values less than or equal to 5 were suppressed by the Philadelphia Public Health Department. We apply the likelihood-based method, the MF, the constant imputation, and the ignoring missing values with an exchangeable, exponential, and CAR covariance model, respectively, to these HIV data. Before applying MF imputation, the data are converted to a matrix that each row is a time point for 19 locations, and each column is a time series for a location. A constant, 2.5, is used for the constant imputation. We use the data at the first 27 time points for training and the data at the last time point for testing.

Figure 6 shows the prediction RMSE and bias from different methods and covariance models. The comparison between different methods in Figure 6 shows a consistent pattern with the simulation results. The likelihood-based and the MF computation methods are the top winners among the four methods. Apparently, the CAR covariance model is more appropriate than the exchangeable and exponential covariance model for this data set. The kriging prediction using the CAR model has much lower RMSE and bias than with other covariance models for all methods. In Figure 6a, the likelihood-based method with the CAR model achieves slightly lower prediction RMSE than the MF, though the results are very similar. Figure 6b shows the likelihood-based and MF have significantly reduced prediction bias compared to simply ignoring missing values. The MF provides a slightly smaller bias than the likelihood method for these data. Constant imputation seems to attain the lowest bias; however, its highest RMSE indicates the skill of predictions with constant imputation is poor in capturing the variability of actual observations. This is because the constant imputation can contaminate the inherent correlation structure of the data, resulting in spurious covariance estimates and consequently leading to poor kriging predictions. The low bias associated with this particular constant is merely by chance.

Figure 5 shows predictions from the likelihood-based method and the MF, both employing the CAR model, and the real values of counts of HIV new diagnosis in Philadelphia. The likelihood and MF predictions are very comparable, but both still tend to slightly overestimate the actual values even if the two methods have largely mitigated the bias caused by the truncated observations as shown in Figure 6b. Figure 7 shows an example of imputation from the likelihood-based method and the MF, both again employing the CAR model. All imputed values are smaller than six, and the imputations from the two methods exhibit only subtle differences. This finding, together with Figure 6b, probably indicate that the likelihood model is a little but not severely misspecified for these data.

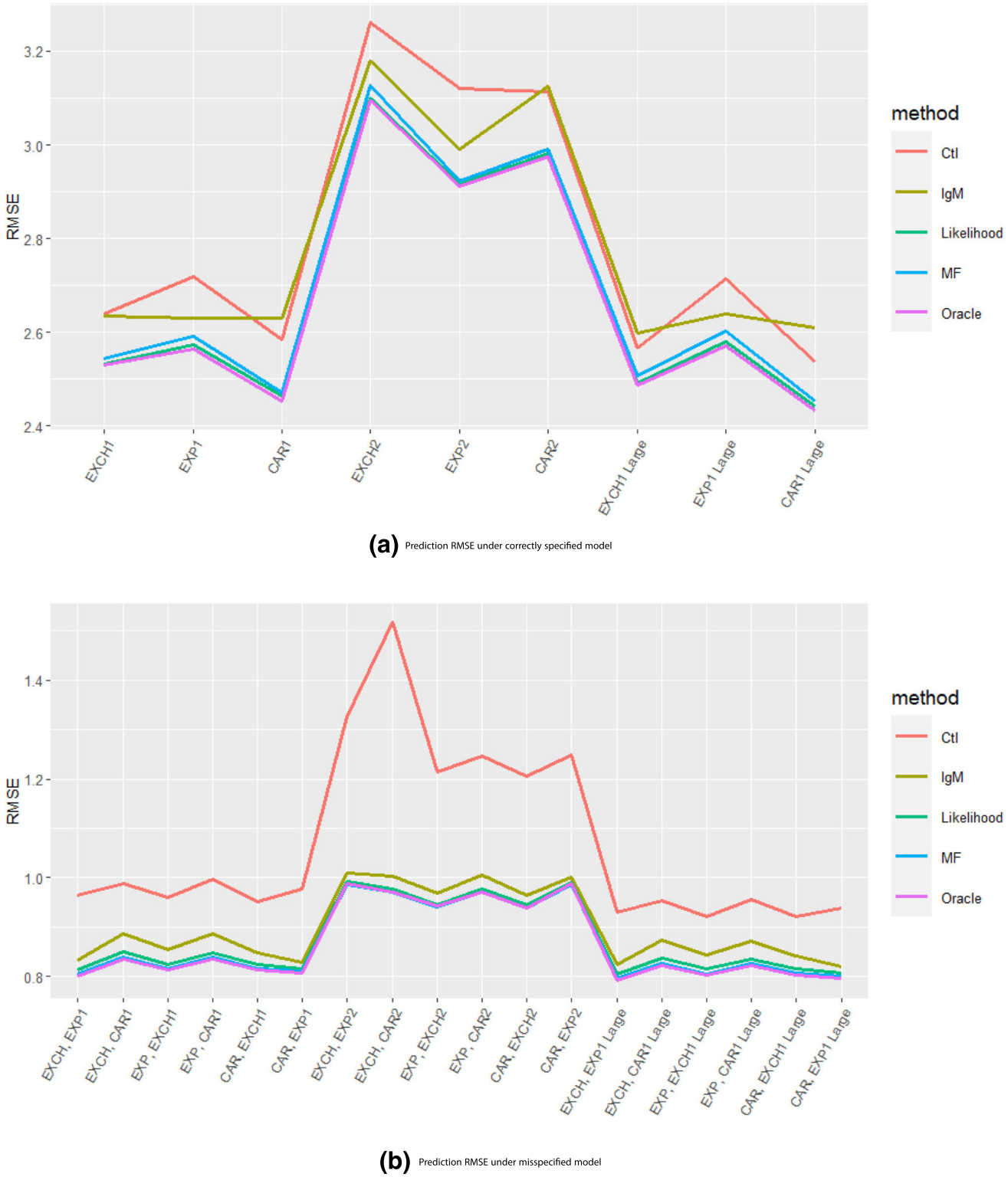


FIGURE 4 Prediction RMSE from five methods under different settings. "Ctl," "IgM," "Likelihood," "MF," and "Oracle," represent constant imputation, ignoring missing values, likelihood based method, MF imputation and oracle prediction, respectively. The x-labels in plot (a) and (b) follow the labels in Figure 3, respectively.

6 | DISCUSSION

Missing values in epidemiology data such as HIV new diagnoses make data analysis challenging, especially when the missingness is not at random and highly unbalanced. We proposed two methods to handle missing values for spatiotemporal data: a traditional likelihood-based method that takes the missing mechanism into account and an adapted MF imputation method. We used simulations and the Philadelphia HIV new diagnosis

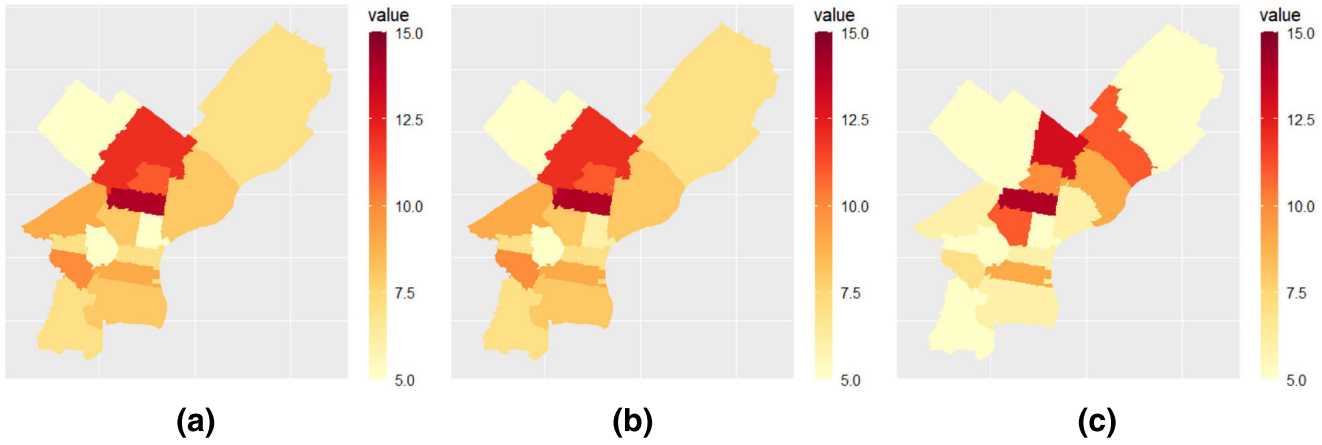


FIGURE 5 Predictions for counts of HIV new diagnosis of 19 Philadelphia zip-code groups in the last quarter of 2015 using (a) Likelihood-based method with CAR model and (b) MF with CAR model. (c) shows the corresponding actual data with suppressed values. Predictions in (a) and (b) that are less than or equal to 5 and missing values in (c) share the same color corresponding to value 5

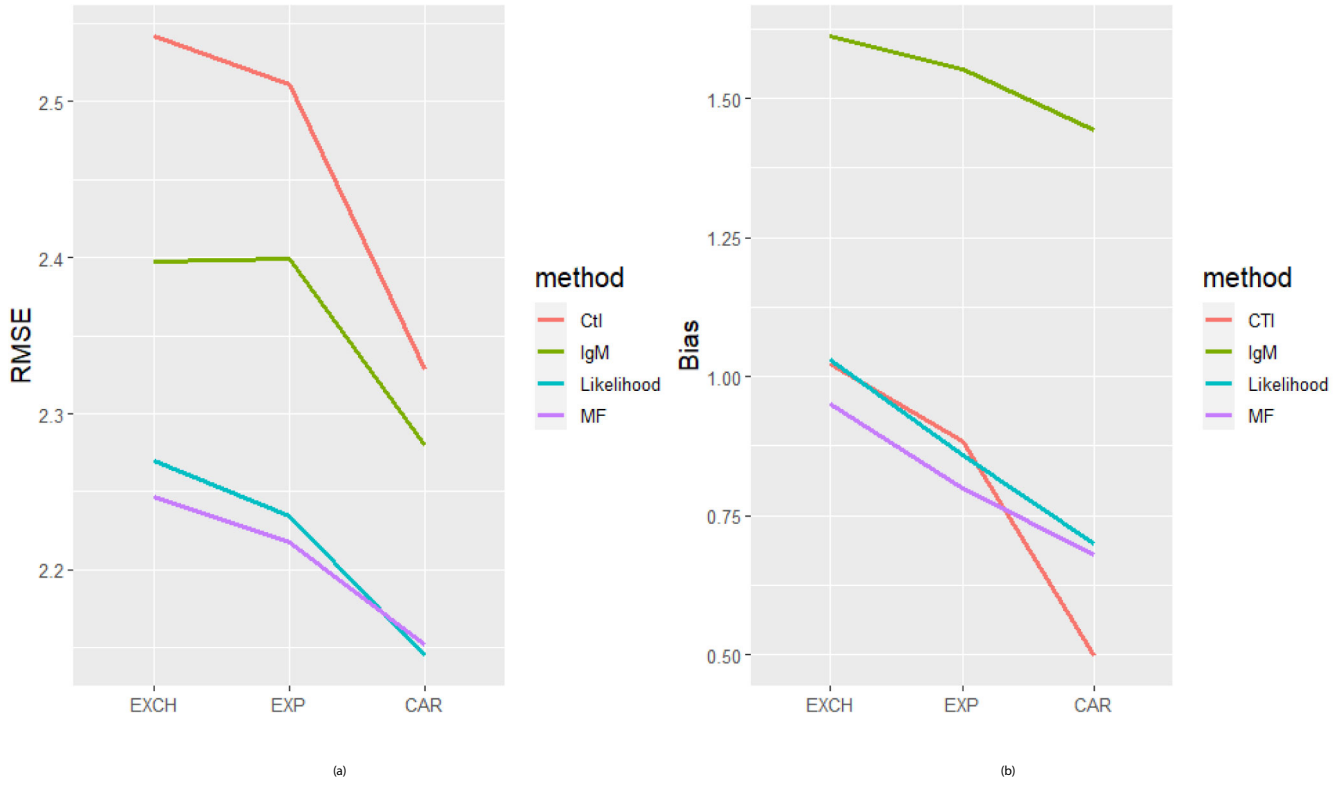


FIGURE 6 (a) Prediction RMSE and (b) Prediction absolute bias of the Philadelphia HIV new diagnosis data based on four methods with three different spatial correlation structures. "MF," "Ctl," "IgG," "Likelihood," and "EXCH," "EXP," and "CAR" are all adopted from Figure 4

counts at the zip code level to study the properties of the two methods and compare them with two baseline methods, imputing the missing values with a preselected constant and ignoring the missing values. Both simulation and real data results show improvement in the predictions and imputations using the proposed methods.

Based on our numerical studies, the likelihood-based method provides more accurate imputations than the MF if the model is correctly specified. If the proposed model deviates from the true model, then the MF imputation, a non-parametric and robust method, seems to provide better-imputed values. However, in terms of prediction, both simulations and real data analysis show that the likelihood-based method is more robust to model misspecification and achieves similar results as MF in all scenarios we consider. In real-world studies, when there is no strong domain knowledge to help specify the correct model, MF imputation method seems a safe choice.

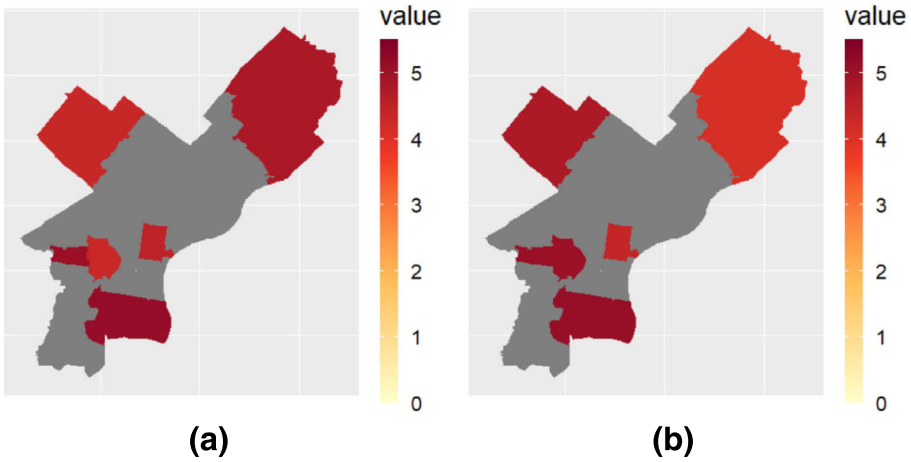


FIGURE 7 Imputed values for counts of HIV new diagnosis of 19 Philadelphia zip-code groups in the last quarter of 2013 using (a) Likelihood-based method with CAR model and (b) MF with CAR model. To highlight the imputation, observed values, values larger than 5, in (a) and (b) are shown in gray.

If the data size is large in the spatial or temporal dimension, the correlation among observations makes the whole data high-dimensional. Due to the curse of dimensionality in sampling, the convergence of MCEM in the likelihood-based method may take a long time, which can be practically impossible. In that case, MF imputation is preferable because the SGD employed in the MF algorithm guarantees a fast convergence rate even with a large sample size.

In this paper, we mainly show derivation and application with the left-censored data. However, both the likelihood-based method and the MF imputation can be generalized to other cases such as interval-censored or right-censored data. They can even handle data that has multiple kinds of censors. Finally, as censored data are common in many health and environment domains, our proposed methods can be applied to other infectious diseases and environmental data modeling and prediction.

ORCID

Tianyi Qu https://orcid.org/0000-0001-7055-7654

REFERENCES

Arnold, R. A., Thomas, A., Waller, L. A., & Conlon, E. M. (1999). Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, *6*, Oxford University Press, pp. 131.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*, *2011*, 27–35.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Camoni, L., Raimondo, M., Regine, V., Salfa, M. C., & Suligoi, B. (2013). Late presenters among persons with a new HIV diagnosis in italy, 2010–2011. *BMC Public Health*, *13*(1), 1–6.

Canales, R. A., Wilson, A. M., Pearce-Walker, J. I., Verhougstraete, M. P., & Reynolds, K. A. (2018). Methods for handling left-censored data in quantitative microbial risk assessment. *Applied and Environmental Microbiology*, *84*(20), e01203.

Chan, M.-S., Lohmann, S., Morales, A., Zhai, C., Ungar, L., Holtgrave, D. R., & Albarracín, D. (2018). An online risk index for the cross-sectional prediction of new HIV chlamydia, and gonorrhea diagnoses across US counties and across years. *AIDS and Behavior*, *22*(7), 2322–2333.

Choi, I., Li, B., & Wang, X. (2013). Nonparametric estimation of spatial and space-time covariance function. *Journal of Agricultural, Biological, and Environmental Statistics*, *18*(4), 611–630.

Cohen Jr, A. C. (1950). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *The Annals of Mathematical Statistics*, *1950*, 557–569.

Cressie, N. (2015). *Statistics for spatial data*: John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 49–73.

Erdman, E. A., Young, L. D., Bernson, D. L., Bauer, C., Chui, K., & Stopka, T. J. (2021). A novel imputation approach for sharing protected public health data. *American Journal of Public Health*, *111*(10), 1830–1838.

Freeman, B. A., Jaro, S., Park, T., Keene, S., Tansey, W., & Reznik, E. (2022). MIRTH: Metabolite imputation via rank-transformation and harmonization. *Genome Biology*, *23*(1), 1–25.

Gleit, A. (1985). Estimation for small normal data sets with detection limits. *Environmental Science & Technology*, *19*(12), 1201–1206.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, *97*(458), 590–600.

Hall, H. I., Song, R., Rhodes, P., Prejean, J., An, Q., Lee, L. M., Karon, J., Brookmeyer, R., Kaplan, E. H., & McKenna, M. T. (2008). Estimation of HIV incidence in the united states. *Jama*, *300*(5), 520–529.

Huang, X.-Y., Li, W., Chen, K., Xiang, X.-H., Pan, R., Li, L., & Cai, W.-X. (2013). Multi-matrices factorization with application to missing sensor data imputation. *Sensors*, *13*(11), 15172–15186.

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*(2), 551–564.

Karon, J. M., Song, R., Brookmeyer, R., Kaplan, E. H., & Hall, H. I. (2008). Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Statistics in Medicine*, *27*(23), 4617–4633.

Kim, J. K., & Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, *106*(493), 157–165.

Kucharska, K., Binkowski, L. J., Zaguła, G., & Dudzik, K. (2022). Spatial, temporal and environmental differences in concentrations of lead in the blood of mute swans from summer and winter sites in Poland. *Science of The Total Environment*, *830*, 154698.

Leacy, F. P., Floyd, S., Yates, T. A., & White, I. R. (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *American Journal of Epidemiology*, *185*(4), 304–315.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence, *Statistical models in epidemiology, the environment, and clinical trials* (pp. 179–191). Springer.

Li, Y., & Ghosh, S. K. (2015). Efficient sampling methods for truncated multivariate normal and Student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice*, *9*(4), 712–732.

Little, R., & Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, *1996*, 1324–1333.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, *112*(518), 484–496.

Mohamed, R. ubaA. M., Brooks, S. C., Tsai, C.-H., Ahmed, T., Rucker, D. F., Ulery, A. L., Pierce, E. M., & Carroll, K. C. (2021). Geostatistical interpolation of streambed hydrologic attributes with addition of left censored data and anisotropy. *Journal of Hydrology*, *599*, 126474.

Moosavi, S. H., Eide, P. W., Eilertsen, I. A., Brunsell, T. H., Berg, K. ajaC. G., Røsok, B. I., Brudvik, K. W., Bjørnbeth, B. A., Guren, M. G., & Nesbakken, A. (2021). De novo transcriptomic subtyping of colorectal cancer liver metastases in the context of tumor heterogeneity. *Genome Medicine*, *13*(1), 1–19.

Murray, J. S., & Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, *111*(516), 1466–1479.

Ndawinz, J. acquesD. A., Costagliola, D., & Supervie, V. (2011). New method for estimating HIV incidence and time from infection to diagnosis using hiv surveillance data: results for france. *Aids*, *25*(15), 1905–1913.

Pakianathan, M., Whittaker, W., Lee, M. J., Avery, J., Green, S., Nathan, B., & Hegazi, A. (2018). Chemsex and new HIV diagnosis in gay, bisexual and other men who have sex with men attending sexual health clinics. *HIV Medicine*, *19*(7), 485–490.

Quick, H. (2019). Peer reviewed: Estimating county-level mortality rates using highly censored data from CDC wonder. *Preventing Chronic Disease*, *16*.

Ranjbar, M., Moradi, P., Azami, M., & Jalili, M. (2015). An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems. *Engineering Applications of Artificial Intelligence*, *46*, 58–66.

Rebouillat, P., Vidal, R., Cravedi, J.-P., Taupier-Letage, B., Debrauwer, L., Gamet-Payrastre, L., Touvier, M., Hercberg, S., Lairon, D., & Baudry, J. (2021). Estimated dietary pesticide exposure from plant-based foods using NMF-derived profiles in a large sample of French adults. *European Journal of Nutrition*, *60*(3), 1475–1488.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sahoo, I., & Hazra, A. (2021). Contamination mapping in Bangladesh using a multivariate spatial Bayesian model for left-censored data. arXiv preprint arXiv: 2106.15730.

Sass, D., Farkhad, B. F., Li, B., Chan, M.-S., & Albarracín, D. (2021). Are spatial models advantageous for predicting county-level HIV epidemiology across the United States? *Spatial and Spatio-Temporal Epidemiology*, *38*, 100436.

Shand, L., & Li, B. (2017). Modeling nonstationarity in space and time. *Biometrics*, *73*(3), 759–768.

Shand, L., Li, B., Park, T., & Albarracín, D. (2018). Spatially varying auto-regressive models for prediction of new human immunodeficiency virus diagnoses. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*(4), 1003.

Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*(5), 499–521.

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, *10*(1), 1–10.

Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2008). Matrix factorization and neighbor based algorithms for the Netflix prize problem. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pp. 267–274.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., & O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, *61*, 119–131.

Wang, H., & Zhou, L. (2017). Random survival forest with space extensions for censored data. *Artificial Intelligence in Medicine*, *79*, 52–61.

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704.

Wei, R., Wang, J., Jia, E., Chen, T., Ni, Y., & Jia, W. (2018). GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Computational Biology*, *14*(1), e1005973.

Yang, J.-M., Peng, Z.-R., & Lin, L. (2021). Real-time spatiotemporal prediction and imputation of traffic status based on LSTM and graph Laplacian regularized matrix factorization. *Transportation Research Part C: Emerging Technologies*, *129*, 103228.

Yeniay, O. (2005). Penalty function methods for constrained optimization with genetic algorithms. *Mathematical and Computational Applications*, *10*(1), 45–56.

Yuan, Y., & Yin, G. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, *66*(1), 105–114.

Zhao, J., & Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*(512), 1577–1590.

Zhou, H., & Hanson, T. (2018). A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. *Journal of the American Statistical Association*, *113*(522), 571–581.

How to cite this article: Qu, T., Li, B., Sally Chan, M., & Albarracin, D. (2023). Bias correction for nonignorable missing counts of areal HIV new diagnosis. *Stat*, *12*(1), e555. https://doi.org/10.1002/sta4.555

APPENDIX A: DERIVATION OF EQUATIONS

A.1 | Derivation of (2.4)

From the relationship between the marginal distribution of $(\boldsymbol{R},\boldsymbol{Z}_O)$ and joint distribution of $(\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{Z}_M)$, the likelihood on the left-hand side of (2.4) can be written as

$$L(\boldsymbol{\theta},\Sigma|\boldsymbol{R},\boldsymbol{Z}_O) = \mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{X},\boldsymbol{\theta},\Sigma) \tag{A1}$$

$$= \int \mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{Z}_M|\boldsymbol{X},\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M. \tag{A2}$$

Then by the chain rule of probability and (2.3), we have

$$\begin{aligned}&\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{Z}_M|\boldsymbol{X},\boldsymbol{\theta},\Sigma)\\ &= \mathrm{pr}(\boldsymbol{R}|\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\theta},\Sigma)\mathrm{pr}(\boldsymbol{Z}_O,\boldsymbol{Z}_M|\boldsymbol{X},\boldsymbol{\theta},\Sigma)\\ &= \prod_{i=1}^{I}\prod_{j=1}^{J}\mathbf{1}(z_{ij}\in\boldsymbol{A})^{r_{ij}}\mathbf{1}(z_{ij}\in\boldsymbol{A}^c)^{1-r_{ij}}\mathrm{pr}(\boldsymbol{Z}_O,\boldsymbol{Z}_M|\boldsymbol{X},\boldsymbol{\theta},\Sigma).\end{aligned}$$

After transforming the indicator function to the domain of integration, (A1) can be written as

$$\begin{aligned}&L(\boldsymbol{\theta},\Sigma|\boldsymbol{R},\boldsymbol{Z}_O)\\ &= \int_{\boldsymbol{Z}_M\in\boldsymbol{A}} \mathrm{pr}(\boldsymbol{Z}_O,\boldsymbol{Z}_M|\boldsymbol{X},\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M.\end{aligned}$$

Finally, by chain rule of probability and model (2.1), the likelihood is

$$\begin{aligned}&L(\boldsymbol{\theta},\Sigma|\boldsymbol{R},\boldsymbol{Z}_O)\\ &= \int_{\boldsymbol{Z}_M\in\boldsymbol{A}} \mathrm{pr}(\boldsymbol{Z}_M|\boldsymbol{Z}_O,\boldsymbol{X},\boldsymbol{\theta},\Sigma)\mathrm{pr}(\boldsymbol{Z}_O|\boldsymbol{X},\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M\\ &= \phi(\boldsymbol{Z}_O;\boldsymbol{\eta}_O(\boldsymbol{\theta},\boldsymbol{X}_O),\Sigma_O)\int_{\boldsymbol{Z}_M\in\boldsymbol{A}} \mathrm{pr}(\boldsymbol{Z}_M|\boldsymbol{Z}_O,\boldsymbol{X},\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M.\end{aligned}$$

A.1.1 | Derivation of (2.6)

By the Bayesian rule, the probability has the following format:

$$\mathrm{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta},\Sigma) = \frac{\mathrm{pr}(\boldsymbol{R}|\boldsymbol{Z}_O,\boldsymbol{Z}_M,\boldsymbol{\theta},\Sigma)\mathrm{pr}(\boldsymbol{Z}_M,\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma)}{\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma)}.$$

From (2.3) and (2.1), the probability can be written as

$$\mathrm{pr}(\boldsymbol{Z}_M|\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{\theta},\Sigma) = \frac{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\prod_{(i,j)\in\mathcal{M}}\mathbf{1}(Z_{i,j}\in(-\infty,\log(5)])}{\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma)}.$$

A.1.2 | Derivation of (2.7)

If we plug (2.6) into (2.5), we have

$$Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right) = \int \frac{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\prod_{(ij)\in\mathcal{M}}\mathbf{1}(Z_{ij}\in(-\infty,\log(5)])}{\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma)}\log\{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\}d\boldsymbol{Z}_M.$$

After transforming the indicator function to the domain of the integration, we have

$$Q\left(\boldsymbol{\theta},\Sigma|\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right) = \frac{\int_{\boldsymbol{Z}_M\in(-\infty,\log(5)]}\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\log\{\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)\}d\boldsymbol{Z}_M}{\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma)}. \tag{A3}$$

For the denominator, the density can be written as an integration of a joint density

$$\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma) = \int \mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O,\boldsymbol{Z}_M|\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M.$$

By the conditional density and the missing mechanism, we can simplify the denominator as

$$\mathrm{pr}(\boldsymbol{R},\boldsymbol{Z}_O|\boldsymbol{\theta},\Sigma) = \int_{\boldsymbol{Z}_M\in(-\infty,\log(5)]}\phi(\boldsymbol{Z}_M,\boldsymbol{Z}_O;\boldsymbol{\theta},\Sigma)d\boldsymbol{Z}_M. \tag{A4}$$

Then we can plug (A4) in (A3) and get (2.7).

APPENDIX B: STOCHASTIC GRADIENT DECENT ALGORITHM FOR MATRIX FACTORIZATION

Algorithm 1 SGD for the MF

initialize parameters P, Q, e and f
for i=1 to N **do** randomly pick a point (i, j) from the data matrix
  if (i, j) is observed **then**
    $res_{ij}^2 = (r_{ij} - \hat{r}_{ij})^2$
    $p'_{ik} = p_{ik} + \alpha(2 \times res_{ij}q_{kj} - \beta p_{ik})$
    $q'_{kj} = q_{kj} + \alpha(2 \times res_{ij}p_{ik} - \beta q_{kj})$
    $e'_i = e_i + \alpha \times (res_{ij} - \theta e_i)$
    $f'_j = f_j + \alpha \times (res_{ij} - \theta f_i)$
  else
    $res_{ij}^2 = (a - \hat{r}_{ij})_+^2 + (\hat{r}_{ij} - b)_+^2$
    $p'_{ik} = p_{ik} + \alpha(2\eta res_{ij}q_{kj} - \beta p_{ik})$
    $q'_{kj} = q_{kj} + \alpha(2\eta res_{ij}p_{ik} - \beta q_{kj})$
    $e'_i = e_i + \alpha \times (\eta res_{ij} - \theta e_i)$
    $f'_j = f_j + \alpha \times (\eta res_{ij} - \theta f_i)$
  end if
end for

APPENDIX C: PARAMETERS IN SIMULATION

TABLE C1 $\boldsymbol{\theta}$ for correctly specified model; θ_i is the approximate average value of the log counts of HIV new diagnosis from zip-code group i.

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
1.9	2	2.1	1.7	2.5	2.5	2.1	1.7	2.3	2.5
θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	
2.2	1.9	2.2	2.1	2.3	1.7	2.3	2.2	2.8	

TABLE C2 λ, ρ and σ for correctly specified model.

	EC1	EXP1	CAR1	EC2	EXP2	CAR2
λ	0.6	0.6	0.6	0.2	0.2	0.2
ρ	0.4	10	0.4	0.3	20	0.2
σ	0.3	0.3	0.3	0.3	0.3	0.3

Note: In the column names, EC is exchangeable spatial model, EXP is exponential model, and CAR is the CAR model. The number 1 represents the stronger correlation setting and 2 the weaker correlation setting.

APPENDIX D: SIMULATION BIAS PLOTS

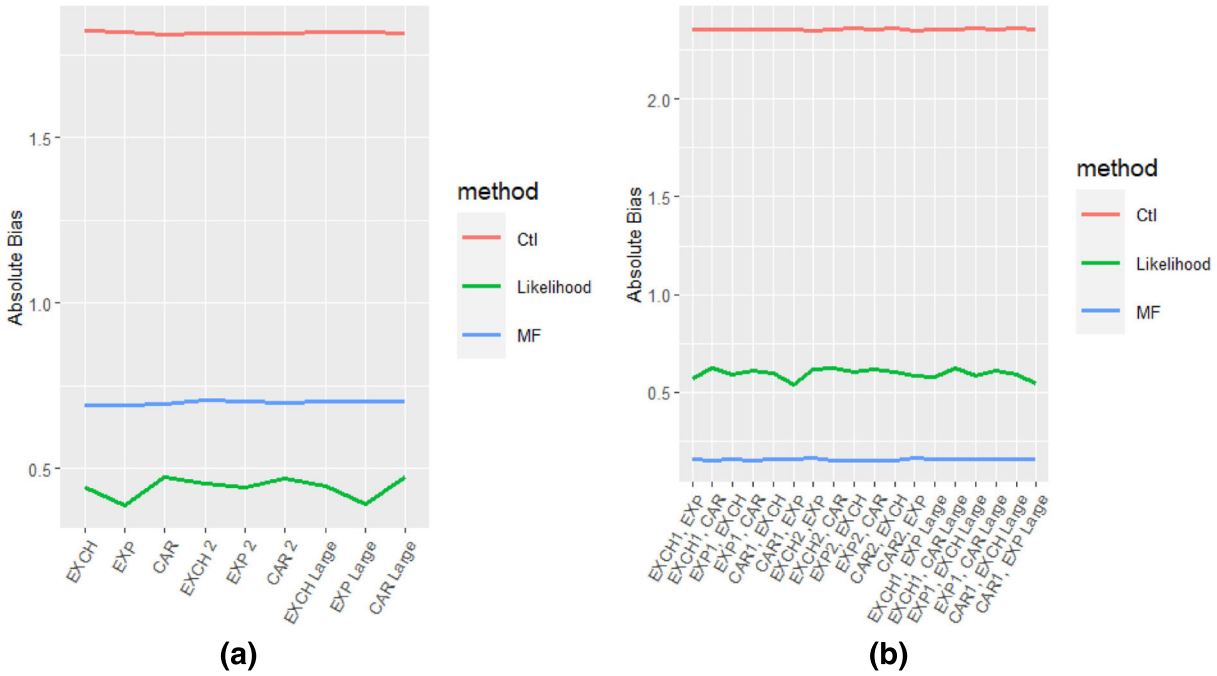
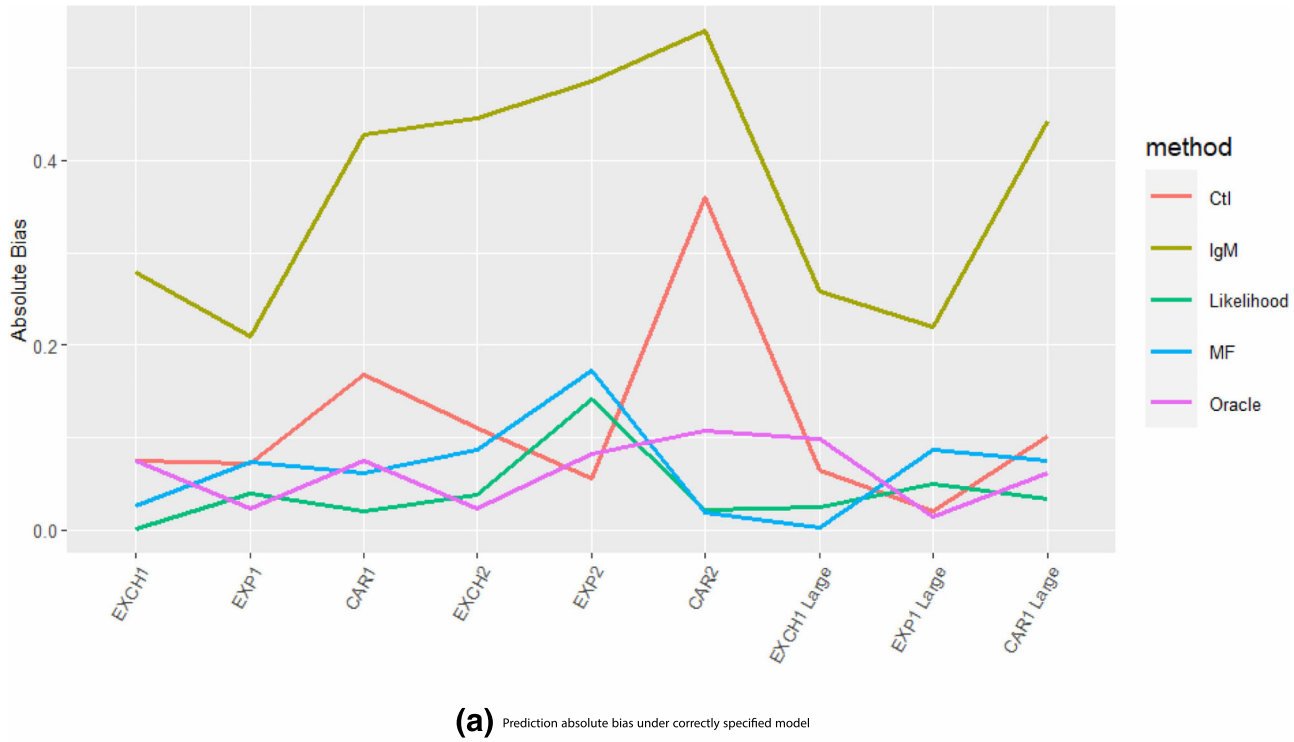
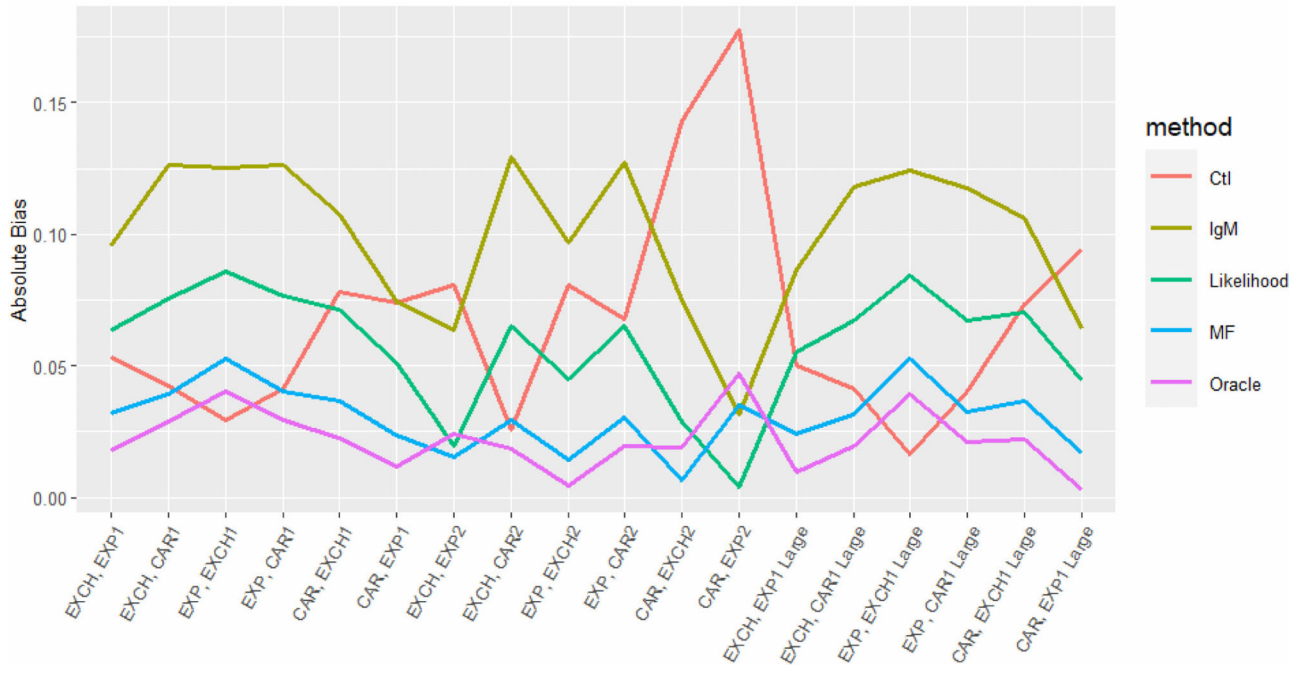


FIGURE D1 Imputation absolute bias over all missing values from the constant imputation (Ctl), likelihood based method and MF imputation under different settings for (a) correctly specified model and (b) misspecified model. The x-labels in plot (a) and (b) follow the labels in Figure 3, respectively.



(a) Prediction absolute bias under correctly specified model



(b) Prediction absolute bias under misspecified model

FIGURE D2 Prediction absolute bias from five methods under different settings. "Ctl," "IgM," "Likelihood," "MF," and "Oracle" represent constant imputation, ignoring missing values, likelihood based method, MF imputation, and oracle prediction, respectively. The x-labels in plot (a) and (b) follow the labels in Figure 3, respectively.