Inferring population HIV viral load from a single HIV clinic's electronic health record

Neal D Goldstein, Justin Jones, Deborah Kahal, Igor Burstyn

Table of Contents

Inferring population HIV viral load from a single HIV clinic's electronic health record

Neal D Goldstein[1]; Justin Jones[1]; Deborah Kahal[2]; Igor Burstyn[1]

[1]Drexel University Philadelphia US
[2]AIDS Care Group Sharon Hill US

Corresponding Author:
Neal D Goldstein
Drexel University
3215 Market St.
Philadelphia
US

Abstract

Background: Population viral load (VL), the most comprehensive measure of the HIV transmission potential, cannot be directly measured due to lack of complete sampling of all people with HIV.

Objective: A given HIV clinic's electronic health record (EHR), a biased sample of this population, may be used to attempt to impute this measure.

Methods: We simulated a population of 10,000 individuals with VL calibrated to surveillance data with a logarithmic mean of 3.6. Three hypothetical EHRs sampled from A) the source population, B) those diagnosed, and C) those retained in care. Our analysis imputed population VL from each EHR using sampling weights followed by Bayesian adjustment. These methods were then tested using patient data from an HIV clinic EHR in Delaware.

Results: Following weighting, the estimates moved in the direction of the population value with correspondingly wider 95% intervals as follows, clinic A: 3.64 (3.29, 4.05), clinic B: 3.65 (3.28, 4.01), and clinic C: 2.38 (2.05, 2.75). Bayesian adjusted weighting further improved the estimate.

Conclusions: These findings suggest that methodological adjustments are ineffective for estimating population VL from a single clinic's EHR without the resource-intensive elucidation of an informative prior.

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
   Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
   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

Inferring population HIV viral load from a single HIV clinic's electronic health record

Neal D. Goldstein[a], Justin Jones[a], Deborah Kahal[b], Igor Burstyn[c]

[a] Department of Epidemiology and Biostatistics, Drexel University Dornsife School of Public Health,

Philadelphia, PA, USA

[b] AIDS Care Group, Sharon Hill, PA, USA

[c] Department of Environmental and Occupational Health, Drexel University Dornsife School of

Public Health, Philadelphia, PA, USA

Correspondence: Neal D. Goldstein, 3215 Market St., Philadelphia, PA, 19104, United States, 1-267-

359-6207, ng338@drexel.edu

Abstract

Background: Population viral load (VL), the most comprehensive measure of the HIV transmission potential, cannot be directly measured due to lack of complete sampling of all people with HIV.

Objective: A given HIV clinic's electronic health record (EHR), a biased sample of this population, may be used to attempt to impute this measure.

Methods: We simulated a population of 10,000 individuals with VL calibrated to surveillance data with a geometric mean of 4,449 copies/mL. Three hypothetical EHRs sampled from A) the source population, B) those diagnosed, and C) those retained in care. Our analysis imputed population VL from each EHR using sampling weights followed by Bayesian adjustment. These methods were then tested using EHR data from an HIV clinic in Delaware.

Results: Following weighting, the estimates moved in the direction of the population value with correspondingly wider 95% intervals as follows, clinic A: 4,364 (1,963, 11,132) copies/mL, clinic B: 4,420 (1,913, 10,199) copies/mL, and clinic C: 242 (113, 563) copies/mL. Bayesian adjusted weighting further improved the estimate.

Conclusions: These findings suggest that methodological adjustments are ineffective for estimating population VL from a single clinic's EHR without the resource-intensive elucidation of an informative prior.

Introduction

There has been increasing interest in using electronic health record (EHR) data as part of public health surveillance efforts [1]. In an interview conducted among local health departments, Comer et al. reported 23 such uses, including incidence or prevalence of infectious and chronic diseases, such as diabetes, hepatitis B and C, asthma, and depression, and uptake of disease prevention programs, including vaccination and HIV testing [2]. Uptake of HIV testing is especially relevant and timely given the 2019 U.S. Department of Health and Human Services' "Ending the HIV Epidemic: A Plan for America" initiative [3]. The plan calls for a 75% reduction in the number of new HIV diagnoses within 5 years, and a 90% reduction within 10 years.

To realize this ambitious goal, health departments monitor data on HIV in their jurisdictions. There are a variety of metrics for doing so including incidence, prevalence, late diagnoses, and viral load (VL), a marker for the success of HIV testing programs and connection to care and treatment. Undetectable VL is the desired outcome in the HIV care continuum because an undetectable VL equates to zero transmission risk, the foundation of treatment as prevention [4]. A hierarchy of aggregated VL measures exist and relate to the natural sampling process that occurs from the source population when individuals are diagnosed (community VL), connected to care (in-care VL), and have VL measures obtained (monitored VL) [5]. The broadest categorization, population VL, is the most comprehensive measure of the HIV transmission potential. However, population VL cannot be directly measured due to lack of complete sampling of the population of people living with HIV (PLWH) as well as lack of complete or recent viral load (VL) data among those diagnosed [5]. Despite its utility and appeal, the measure has notable challenges, including population selection, varying definitions and calculations, and complete and accurate surveillance [6]. These issues may have led to the decline in its use following its introduction in 2009. Nevertheless, population viral load – if quantifiable – is a useful latent measure of transmission potential and quality of HIV care and treatment in a specific geographic area. Even a biased measure can be useful if it can be

calibrated to a less biased or an unbiased measure. For example, one contemporary paper using data from the 2010s investigated community VL and HIV incidence in South Carolina and found that community VL disparities mirrored disparities in HIV access to care for non-prioritized groups including women, rural populations, and heterosexual transmission [7].

Absent complete (or a representative random) sampling of a population of PLWH, one may turn to EHRs from various clinics to estimate population VL. A given health department might wish to know the distribution of VL among PLWH in its jurisdiction but only have a single HIV care program that serves the community. As such, the ability to estimate population VL from a single EHR may be of value. In fact, researchers have previously demonstrated how EHR data can improve the accuracy of HIV surveillance programs [8]. However, use of EHRs for these purposes face methodological challenges, including ambiguous catchment [9]. A given EHR can be expected to over or under-sample with respect to characteristics of PLWH (e.g., health, income, race, age, distance to clinic, etc.). We sought to investigate the feasibility of imputing population VL from a single EHR and under what conditions this may be possible.

Materials and Methods

Creation of the synthetic dataset and clinics

To establish the feasibility of recovering the true population VL from a single clinic's EHR, we would need both clinic-level VL EHR data as well as the VL from the source population, data which are difficult to obtain as this would require measuring VL among those unaware of their HIV status as well as those not engaged in care. In lieu of this, we created a hypothetical synthetic source population: this population can be considered a large urban area in the U.S. with a population size of 1,000,000 people and 1% HIV seropositivity, or 10,000 PLWH. We defined three demographic strata for the population: age (<35 years, 35-44 years, 45-54 years, >54 years), gender (male, female), and race/ethnicity (non-Hispanic White, non-Hispanic Black or African American, Hispanic or Latino).

These categories were not meant to be inclusive of all risk groups but rather commonly reported groups for calibrating VL.

The demography of PLWH was randomly sampled from a uniform distribution with probabilities informed from the CDC 2020 HIV Surveillance Report [10]. Specifically, approximately 75% of the population was set to male and 25% to female; 18% <35 years, 19% 35-44 years, 24% 45-54 years, and 39% >54 years; and 33% White, 45% Black or African American, and 23% Hispanic or Latino. VL was randomly sampled from a log-normal distribution with a $\log_{10}$ geometric mean of 3.65 (4,449 copies/mL) and a $\log_{10}$ standard deviation of 1.2. The mean was informed from the measured community VL from the San Francisco, California HIV/AIDS Case Surveillance System for 2005-2008 [11], and the standard deviation was informed from CDC's guidance document on community VL [5]. VL was adjusted jointly across the demographic strata by multiplying the VL by a randomly sampled probability obtained from a normal distribution with the following means and accompanying standard deviation of 10%: -1% male and +18% female; +21% <35 years, -10% 35-44 years, -26% 45-54 years, and -26% >54 years; and -10% White, +13% Black or African American, and +15% Hispanic or Latino. These adjustments were informed from differences observed in VL in the San Francisco surveillance data [11].

To simulate the HIV care continuum from this source population, we set approximately 10% of the population as unaware of their HIV status. This group was more likely to be younger, male, and Black or African American based on a study of concurrent HIV and AIDS diagnosis in San Francisco [12]. Among those aware of their status, we created an "in care" group where approximately 72% of those in care would be virally suppressed (<200 copies/mL), mirroring the 2021 San Francisco HIV epidemiology annual report [13], although we stress that our primary intention is not to replicate San Francisco surveillance data but rather create a hypothetical urban population. Sampling the "in care" group in this manner resulted in an averaged 20% of the aware group also being in care.

Finally, to isolate the effects of various sampling mechanisms, we created three HIV clinics with differing catchments. Clinic A sampled directly from the source population, Clinic B sampled from those aware of their HIV status, and Clinic C sampled from those in care. Each clinic contained 250 active patients oversampled by male sex, White race, and 45 years of age or older. The demographic composition of each clinic was set to reflect observed patterns of retention in HIV care [14] and to yield an EHR where the mean VL differed from the source population. We created 1,000 versions of each clinic to account for random variability.

Creation of the catchment sampling weights and weighted analysis

Let K be the size of the source population, P be the number of people aware of their HIV status, N be the number of people in care, and S be number of patients in care at a clinic. We can estimate the catchment sampling weight by Eq. 1.

Eq 1. $W = 1 / Beta((S + 1), (N + 1 - S))$

In this equation, N arises from $Binomial(N/K, K)$ where N/K is the prevalence of PLWH and in care in the source population. Weights are calculated per the demographic strata enumerated earlier that related to a clinic's catchment (i.e., race, age, and gender) such that P, N, S, and W are all calculated separately for each stratum. The final sample weight is obtained for each person by multiplying the corresponding stratum-specific weights.

To allow for the possibility of weight misspecification when they are not estimated appropriately, for example due to an ambiguous catchment, we transformed W as outlined in Eq. 2.

Eq 2. $P_{biased} = \log(P/(1-P)) + b*\log(VL)$

In this equation, P is the inverse of W, that is, the individual selection probability of being in the clinic, and consequently the inverse of P_{biased} is the misspecified (biased) weights. The coefficient 'b' is the bias factor and was set to 0.1, a conservative starting point that would still meaningfully shift the weights. Under Eq. 2, a positive bias factor demonstrates the scenario whereby individuals with higher VLs are less likely to be sampled in the clinic, but, unbeknownst to the researchers, the catchment model does not identify them as such. Consequently, this bias factor down-weighted their contribution in the weighted analysis by a factor of 0.1, when they should have been up-weighted. Larger bias factors would create greater weight misspecification, albeit with the same conclusions.

We simulated 1,000 of the unbiased and biased weights per participant and then calculated the population geometric mean (GM) VL for each clinic, where   GM = exp(mean(log$_{10}$(VL))). We also calculated the unweighted GM and took the root mean squared error (RMSE) between the weighted and unweighted measures. The final calculations are thus based on the 1,000 weights for each of the 1,000 clinic A's, 1,000 clinic B's, and 1,000 clinic C's. Our target estimand was the median and 95% interval of each clinic's GM distribution.

Post-weighting Bayesian adjustment

Following the weighted analysis, we conducted a Bayesian analysis with the expectation that this will further improve our ability to impute the population VL from a given clinic. This approach is analogous to that taken by others who treated weighted observations as "data" that enters the likelihood part of the Bayesian computation [15]. For this analysis, we assumed the true mean and variance were unknown and specified a Normal-Gamma conjugate prior, although since our focus is only on the posterior mean, the calculations become simplified. The prior mean (μ_0) was informed by the San Francisco HIV/AIDS Case Surveillance System, namely log$_{10}$ GM VL of 3.65. As a starting point for the prior sample size (n_0), we took the perspective of a clinic's population's VL measured at a previous timepoint, i.e., available before the observed VL data used in the weighted analysis. For

example, one might posit that such data were collected immediately upon diagnosis as opposed to routine monitoring during antiretroviral therapy. Following our weighted analysis, these observed measurements have a logarithmic mean of $\acute{x}_w$ and effective sample size, $n_w = \Sigma W$ for each of the 1,000 clinic samples. The posterior logarithmic mean of the population VL (μ_n) conditional on posterior variance is specified in Eq 4.

Eq 4. $\mu_n = ((n_0 \times \mu_0) + n_w \times \acute{x}_w) / (n_0 + n_w)$

To examine the influence of the prior sample size, we operationalized n_0 three ways: $0.25 \times n_w$, $0.5 \times n_w$, and $2 \times n_w$. Additionally, to reflect the earlier scenario of the prior data collected upon diagnosis, we conducted a final analysis for clinic C where μ_0 was informed from clinic B's weighted mean and the more conservative $n_0 = 0.25 * n_w$. As before, these calculations were performed for each of the 1,000 clinic weights for each of the 1,000 clinics.

Real-world clinic dataset

As an applied demonstration of our methods, we obtained the most recently available VL on active patients retained in care for HIV at the Holloway Community Program at ChristianaCare (Wilmington, DE). Patients' age, race, and gender were coded using the same categories defined earlier for our synthetic population. Denominators needed for the catchment model were obtained from U.S. Census Bureau 2021 American Community Survey [16] population sizes for Delaware (the presumed catchment of the Holloway program) and the Delaware Integrated HIV Prevention and Care Plan for 2022-2026 that includes statewide surveillance data as of 2019 [17]. Using the procedures outlined earlier, we estimated the population VL from the clinic data, however as we did not have access to historic unbiased estimates of VL for this jurisdiction, we used the prior as described in our synthetic analysis. To further acknowledge uncertainty in the prior mean (μ_0), we

conducted a sensitivity analysis with μ_0 modified in three ways: $0.25\times\mu_0$, $0.5\times\mu_0$, and $2\times\mu_0$, and repeated the Bayesian adjustment across the three prior sample sizes.

All analyses were performed in R version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria). Analytic codes are available for download from [18]. HIV VL point estimates and 95% intervals are presented on a linear scale in the main text and a logarithmic scale in the multimedia appendix.

Results

Synthetic Population and Clinics

Each clinic was approximately 93% male; 4% <35 years, 5% 35-44 years, 40% 45-54 years, and 51% >54 years; and 81% White, 13% Black or African American, and 6% Hispanic or Latino (Table 1).

Table 1. Characteristics of the synthetic population and clinics as well as the real-world cohort from the Holloway Community Program at ChristianaCare (Wilmington, DE). The three synthetic clinic electronic health records (n=250 per clinic) sampled from a source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Synthetic population results given as the median values from 1,000 hypothetical clinics.
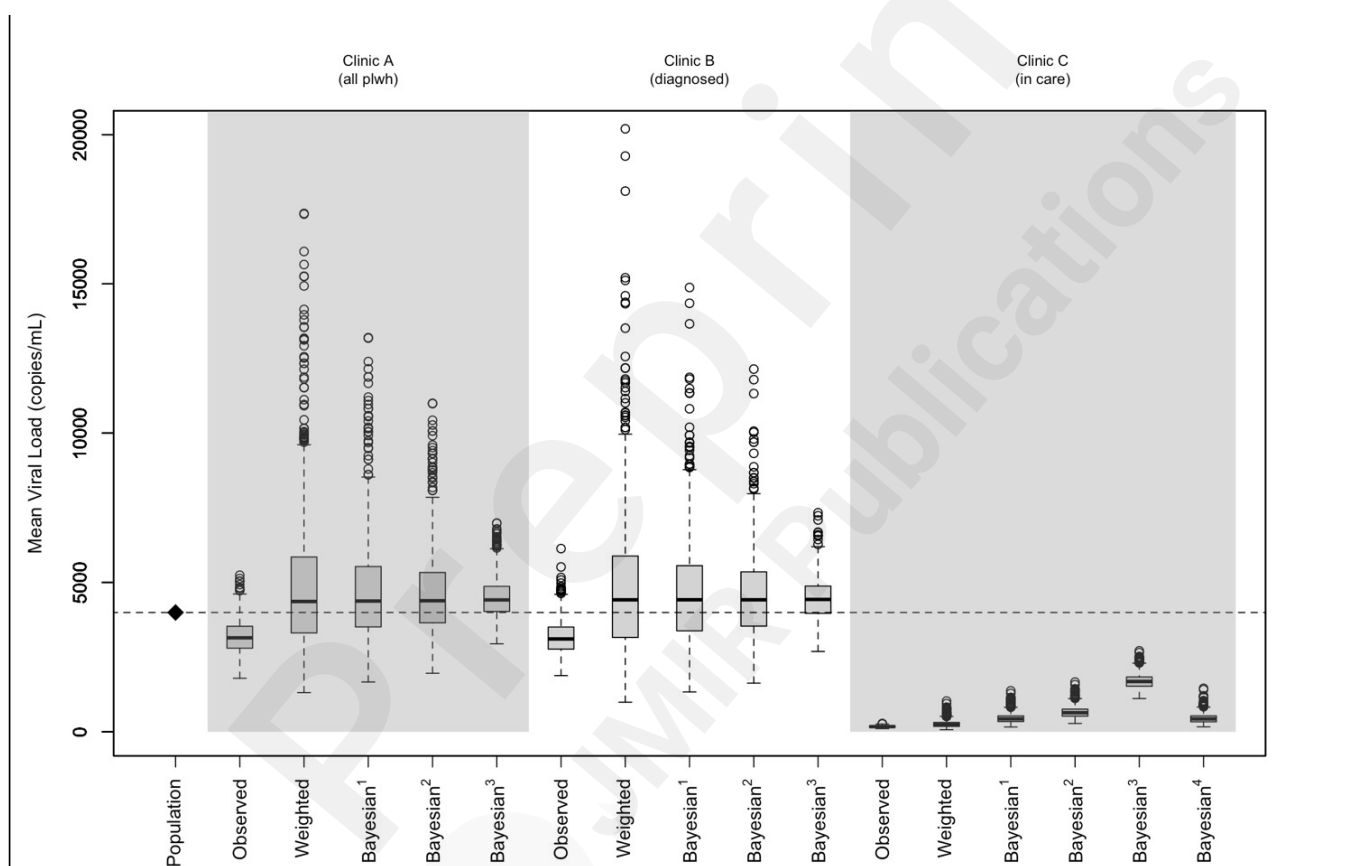
Characteristic	Synthetic EHRs				Real-world EHR
	Population (n=10,000)	Clinic A (n=250)	Clinic B (n=250)	Clinic C (n=250)	Holloway (n=1,807)
Age, n (%)					
<35 years	1,817 (18.2)	12 (4.8)	10 (4)	10 (4)	278 (15.4)
35-44 years	1,819 (18.2)	12 (4.8)	12 (4.8)	15 (6)	299 (16.5)
45-54 years	2,727 (27.3)	97 (38.8)	98 (38.2)	96 (38.4)	332 (18.4)
>54 years	3,634 (36.3)	129 (51.6)	130 (52)	129 (51.6)	898 (49.7)

Gender, n (%)					
Female	2,497 (25.0)	17 (6.8)	17 (6.8)	19 (7.6)	558 (30.9)
Male	7,503 (75.0)	233 (93.2)	233 (93.2)	213 (92.4)	1,249 (69.1)
Race/ethnicity, n (%)					
Non-Hispanic Black or African American	4,446 (44.5)	30 (12)	27 (10.8)	32.5 (13)	1,128 (62.4)
Non-Hispanic White	3,332 (33.3)	205 (82)	207 (82.8)	200 (80)	514 (28.4)
Hispanic or Latino	2,220 (22.2)	15 (6)	15 (6)	18 (7.2)	165 (9.1)
Geometric mean **viral load**, copies/mL	3,996	3,147	3,108	173	41
Log_{10} geometric mean **viral load**	3.60	3.50	3.49	2.24	1.61

Figure 1 contrasts the observed, weighted, and Bayesian adjusted VLs comparing the clinics to the population (Multimedia Appendix 1 for logarithmic results). Across the 1,000 simulations, the median GM population VL was 3,996 (95% interval: 3,780, 4,214) copies/mL. For each clinic A, B, and C, respectively, the median GM VL point estimates and 95% intervals were: 3,147 (2,294, 4,301), 3,108 (2,216, 4,383), and 173 (123, 240) copies/mL. Following weighting, the estimates moved in the direction of the population value with correspondingly wider 95% intervals as follows, clinic A: 4,364 (1,963, 11,132) copies/mL, clinic B: 4,420 (1,913, 10,199) copies/mL, and clinic C: 242 (113, 563) copies/mL. Bayesian adjustment resulted in a shrinking of intervals, depending on the prior sample size, where the large sample size resulted in tighter intervals, and clinic C had a notable shift in point estimates towards the population mean. With a 25% of the clinic prior sample size the posterior estimates were 433 (236, 851) copies/mL, with a 50% of the clinic prior sample size the posterior estimates were 639 (384, 1,120) copies/mL, and with a 200% of the clinic prior sample size the posterior estimates were 1,685 (1,307, 2,231) copies/mL. When using the weighted clinic B estimates to inform the prior for clinic C, we also noted an improvement in estimating the population mean with posterior estimates of 432 (230, 889) copies/mL.

Figure 1. Comparison of the distribution of the geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people

living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. [1] Prior sample size of 25% of the weighted clinic sample size. [2] Prior sample size of 50% of the weighted clinic sample size. [3] Prior sample size of 200% of the weighted clinic sample size. [4] Prior mean informed from weighted clinic B estimates.
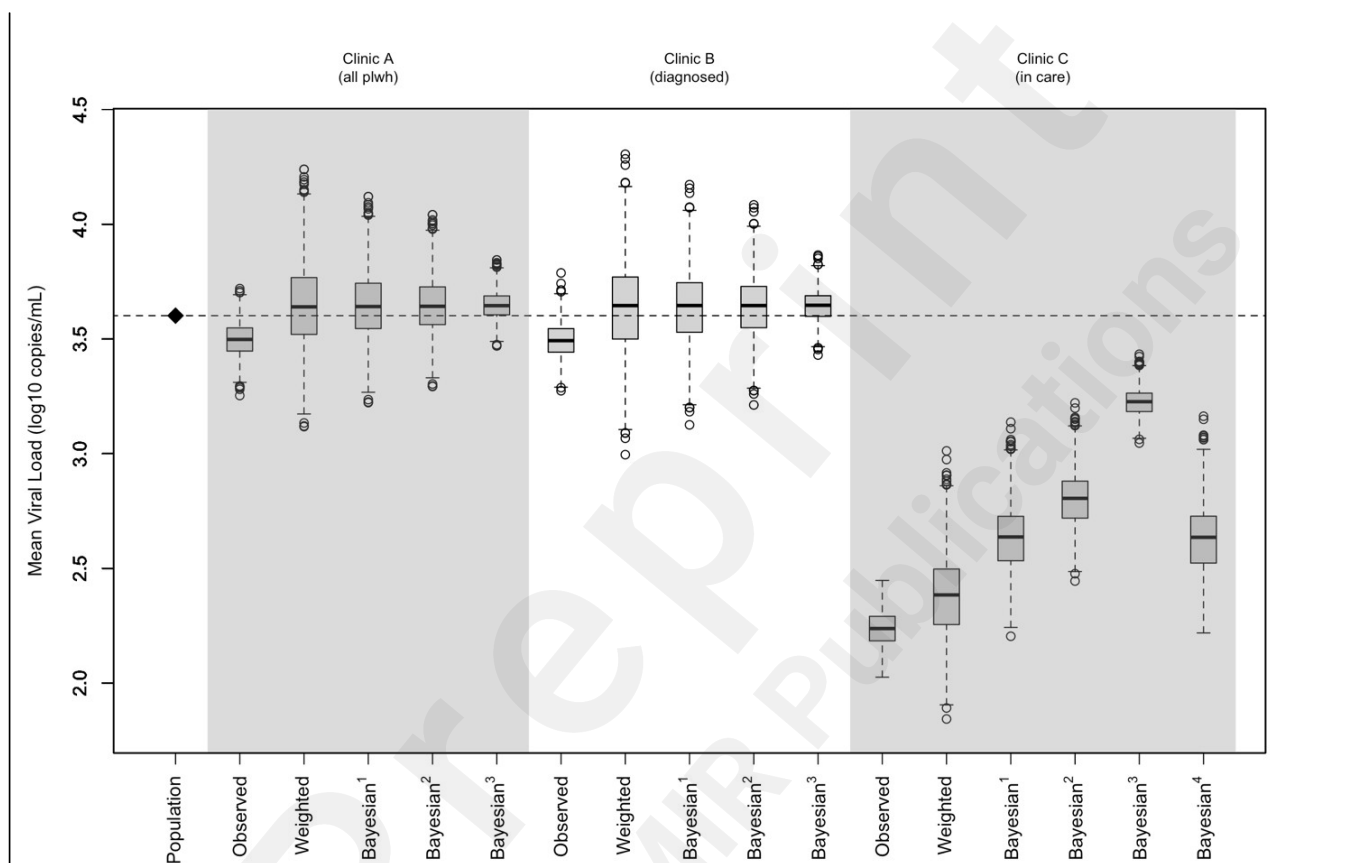

Multimedia Appendix 1. Comparison of the distribution of the logarithmic geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population

based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled

by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics

each with 1,000 sampling weight adjustments. [1] Prior sample size of 25% of the weighted clinic

sample size. [2] Prior sample size of 50% of the weighted clinic sample size. [3] Prior sample size of

200% of the weighted clinic sample size. [4] Prior mean informed from weighted clinic B estimates.
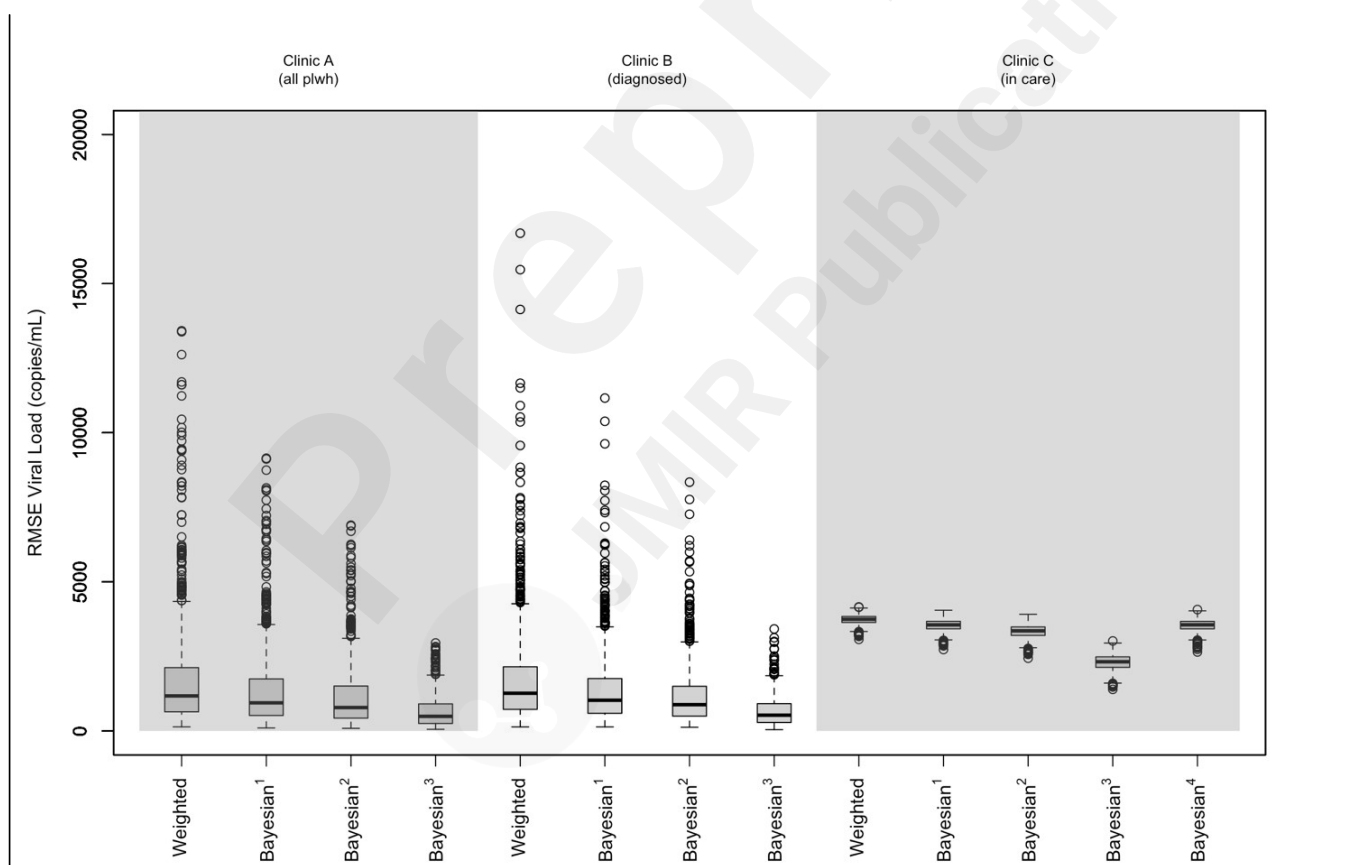

Figure 2 depicts the averaged RMSE for each clinic for each weighting strategy (Multimedia

Appendix 2 for logarithmic results). RMSE was greatest in the purely weighted analyses, with

median errors and 95% intervals for each clinic as follows, clinic A: 1,174 (288, 7,261) copies/mL,

clinic B: 1,265 (261, 6,369) copies/mL, and clinic C: 3,745 (3,385, 4,018) copies/mL. RMSE was

lowest in the Bayesian analysis that followed weighting with the larger prior sample size, as follows:

clinic A: 490 (92, 2,026) copies/mL, clinic B: 528 (96, 1,884) copies/mL, and clinic C: 2,319 (1,773,

2,747) copies/mL.

Figure 2. Comparison of the root mean squared error (RMSE) of the geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. [1] Prior sample size of 25% of the weighted clinic sample size. [2] Prior sample size of 50% of the weighted clinic sample size. [3] Prior sample size of 200% of the weighted clinic sample size. [4] Prior mean informed from weighted clinic B estimates.

Multimedia Appendix 2. Comparison of the root mean squared error (RMSE) of the logarithmic

geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. [1] Prior sample size of 25% of the weighted clinic sample size. [2] Prior sample size of 50% of the weighted clinic sample size. [3] Prior sample size of 200% of the weighted clinic sample size. [4] Prior mean informed from weighted clinic B estimates.
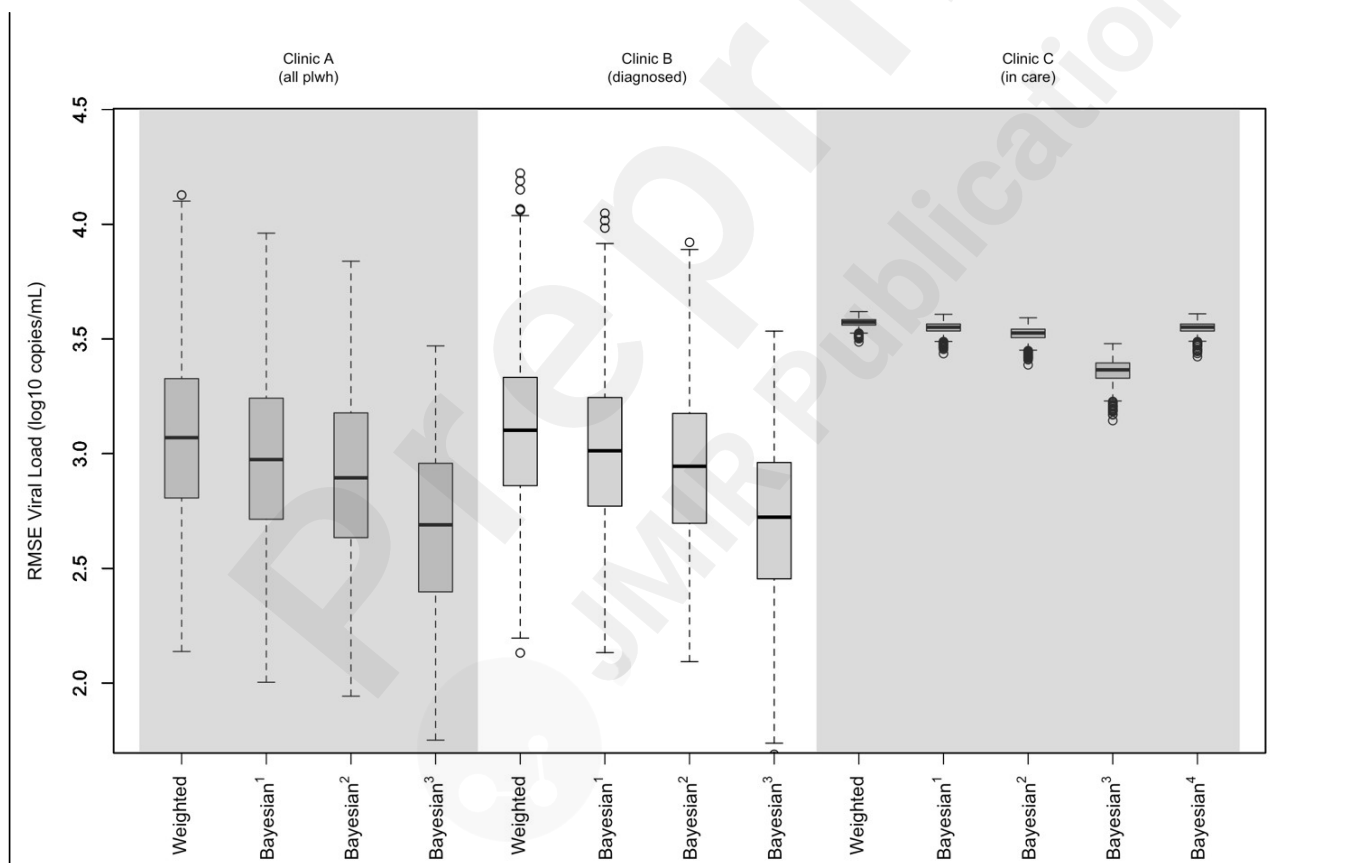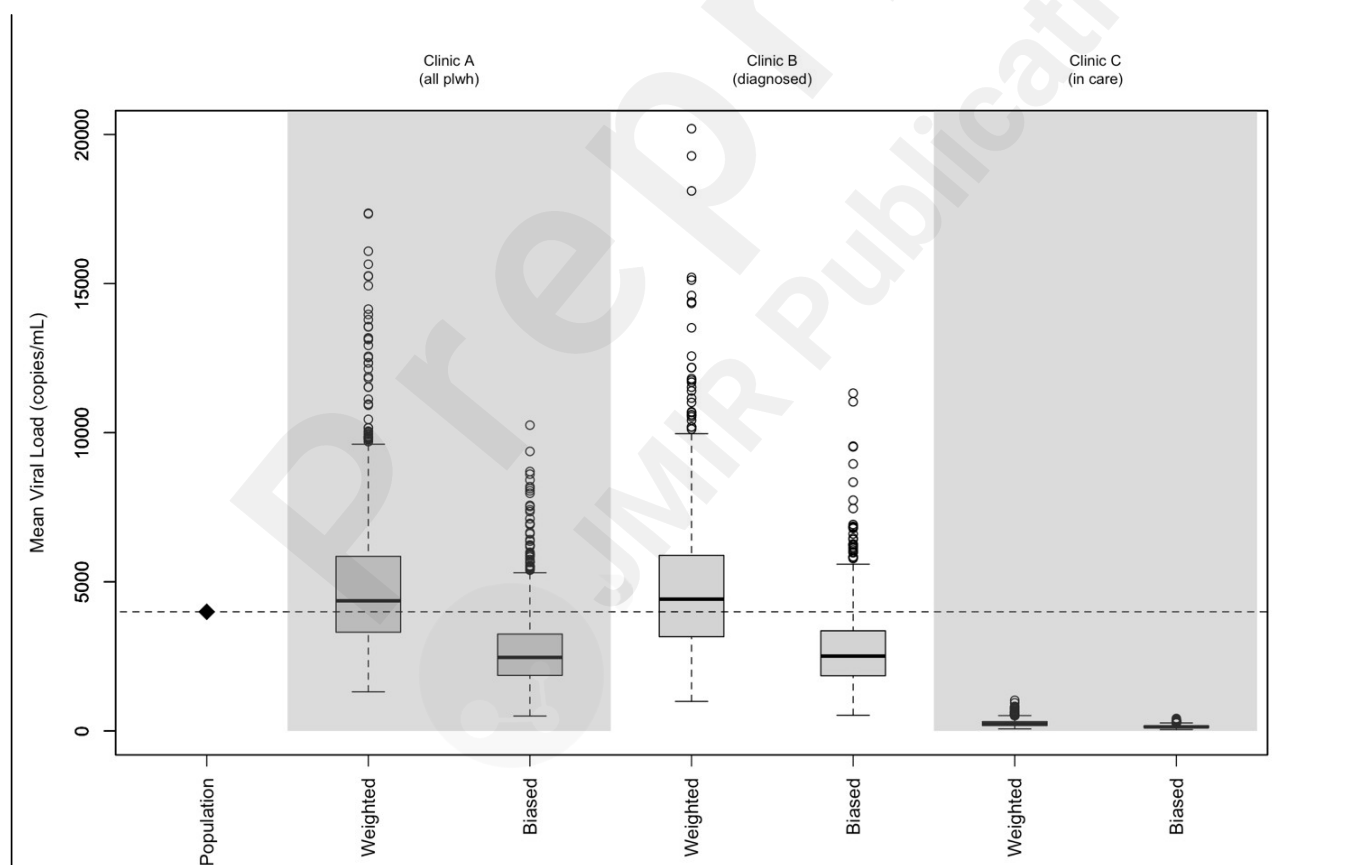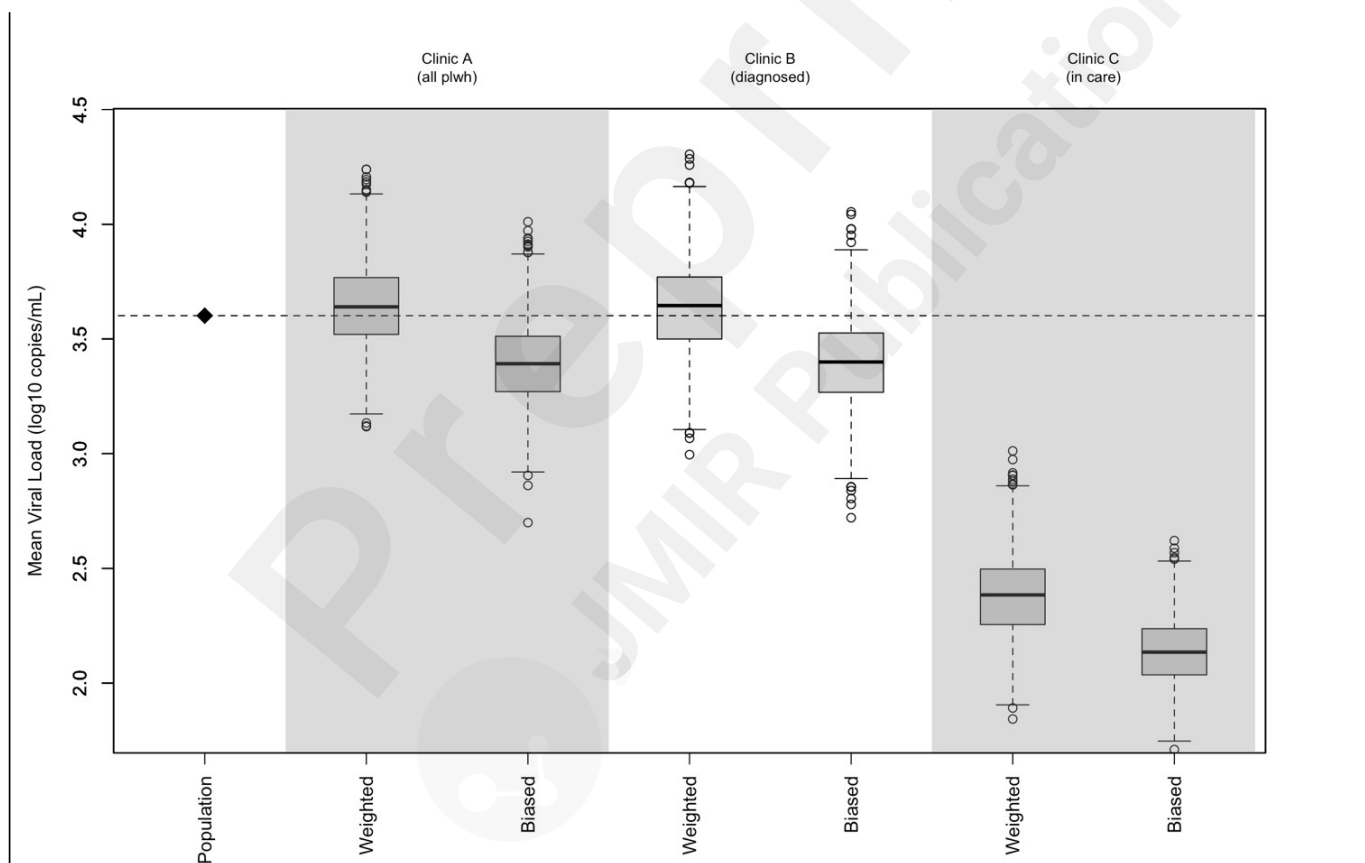

Figure 3 shows the impact of the weight misspecifications (Multimedia Appendix 3 for logarithmic results). As expected, the biased weight resulted in an underestimation of the true population GM VL.

Figure 3. Comparison of weight misspecification in the weighted geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic.

Multimedia Appendix 3. Comparison of weight misspecification in the weighted logarithmic

geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic.

Holloway Community Program Clinic

The 2021 population in Delaware was 1,003,384. For each demographic stratum, the populations were: age (<35 years= 420,844, 35-44 years=122,088, 45-54 years=115,300, >54
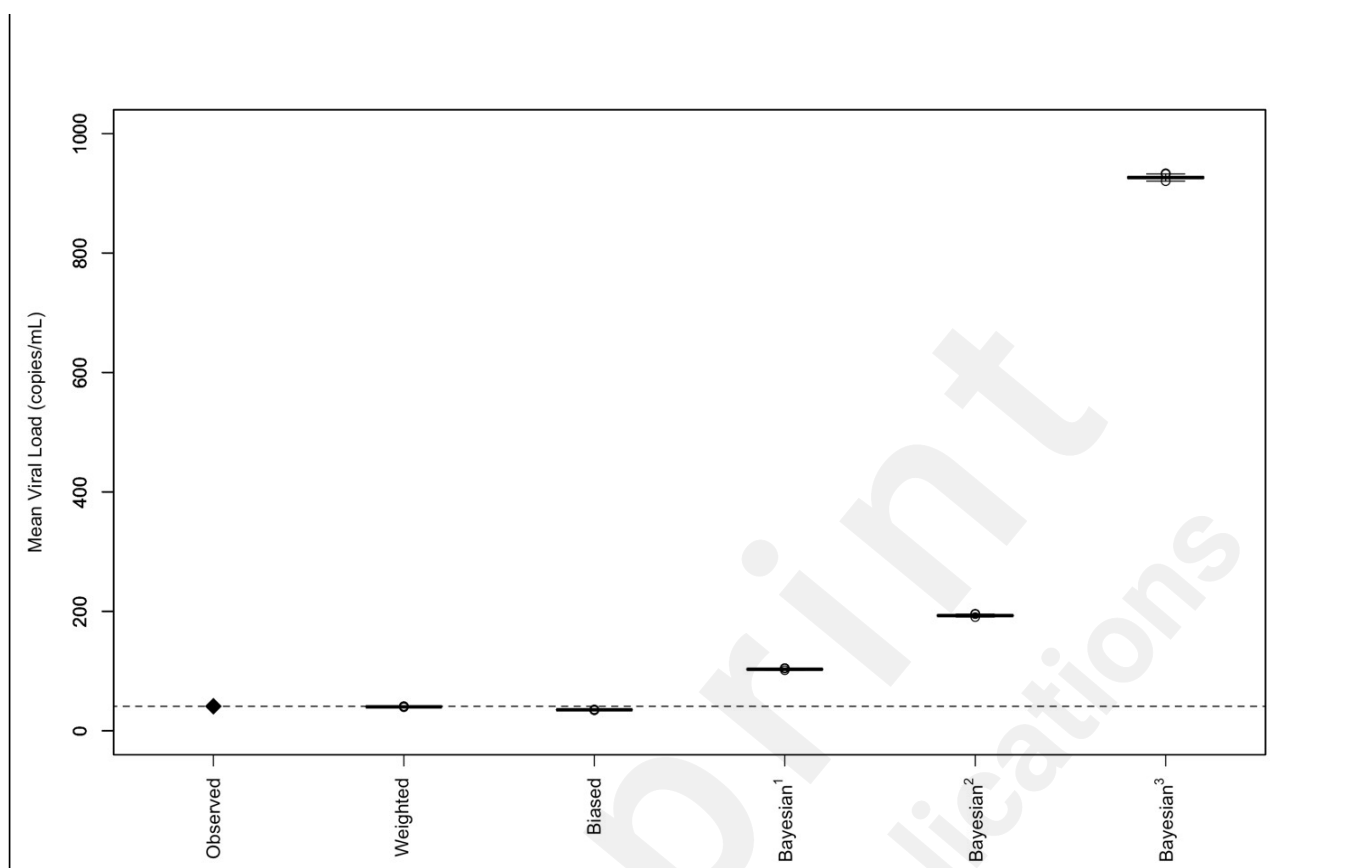
years=345,152), gender (male=485,908, female=517,476), and race/ethnicity (non-Hispanic White=595,212, non-Hispanic Black or African American=205,217, Hispanic or Latino=101,213, other=101,742). As of 2019, there were an estimated 3,841 PLWH; 2,984 were in care and 857 were not in care. For each demographic stratum among those in care, the populations were: age (<35 years= 394, 35-44 years=432, 45-54 years=703, >54 years=1,455), gender (male=2,125, female=859), and race/ethnicity (non-Hispanic White=958, non-Hispanic Black or African American=1,725, Hispanic or Latino=222, other=79).

There were 1,807 active patients in the Holloway Community Program with a resulted VL test as of the date of EHR data extraction. The GM VL of the clinic was 41 and the geometric standard deviation was 190,261 copies/mL; 1,656 (92%) were virally suppressed (<200 copies/mL). Additional characteristics may be found in Table 1.
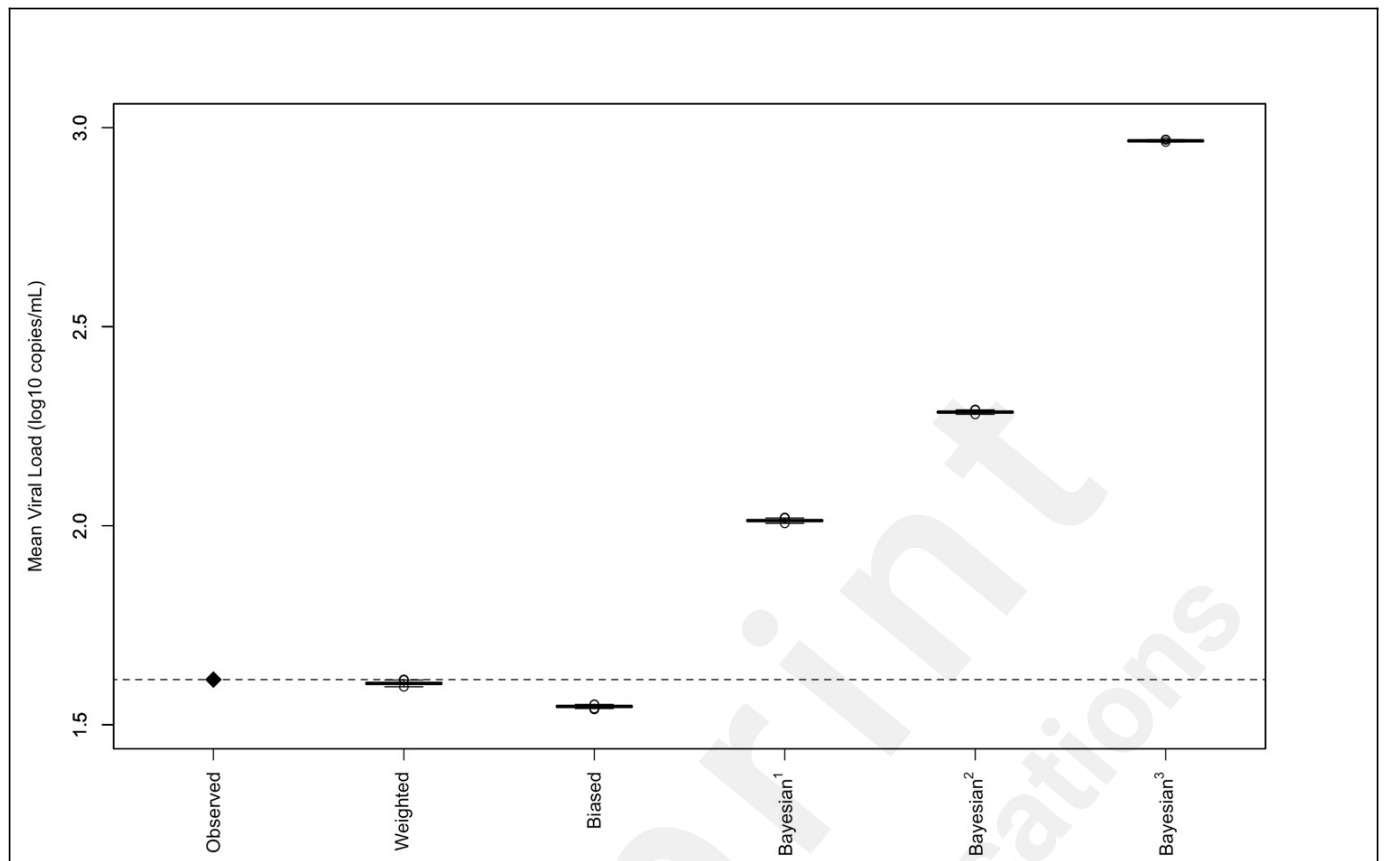
Figure 4 presents the inferred population VL measure from the clinic's EHR (Multimedia Appendix 4 for logarithmic results). The weighting-only adjustment had negligible impact compared to the unweighted estimate, while the biased weights shifted the estimates slightly lower to a median of 35 copies/mL. Meanwhile, the Bayesian adjustment moved the weighted estimate from 40 copies/mL to a median of 103 copies/mL with a 25% prior sample size, to 193 copies/mL with the 50% prior sample size, and to 926 copies/mL with the 200% prior sample size. Results were sensitive to the assumption about the informative prior's mean (Multimedia Appendix 5).

Figure 4. Inferred population geometric mean HIV viral load (VL) for Delaware based on active patients retained in care at the Holloway Community Program at ChristianaCare (Wilmington, DE). Results represent 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic. [1] Prior sample size of 25% of the weighted clinic sample size. [2] Prior sample size of 50%

of the weighted clinic sample size. [3] Prior sample size of 200% of the weighted clinic sample size.

Multimedia Appendix 4. Inferred population geometric mean HIV viral load (VL) for Delaware based on active patients retained in care at the Holloway Community Program at ChristianaCare (Wilmington, DE). Results represent 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic. [1] Prior sample size of 25% of the weighted clinic sample size. [2] Prior sample size of 50% of the weighted clinic sample size. [3] Prior sample size of 200% of the weighted clinic sample size.

Multimedia Appendix 5. Sensitivity analysis of inferred population geometric mean HIV viral load (VL) for Delaware based on active patients retained in care at the Holloway Community Program at ChristianaCare (Wilmington, DE). Results represent 1,000 Bayesian sampling weight adjustments. Sensitivity analysis compared three alternate specifications of the prior mean for VL: 25%, 50%, and 200% of the original specification (m0). Prior sample size was varied three ways: 25%, 50%, and 200% of the weighted clinic sample size.
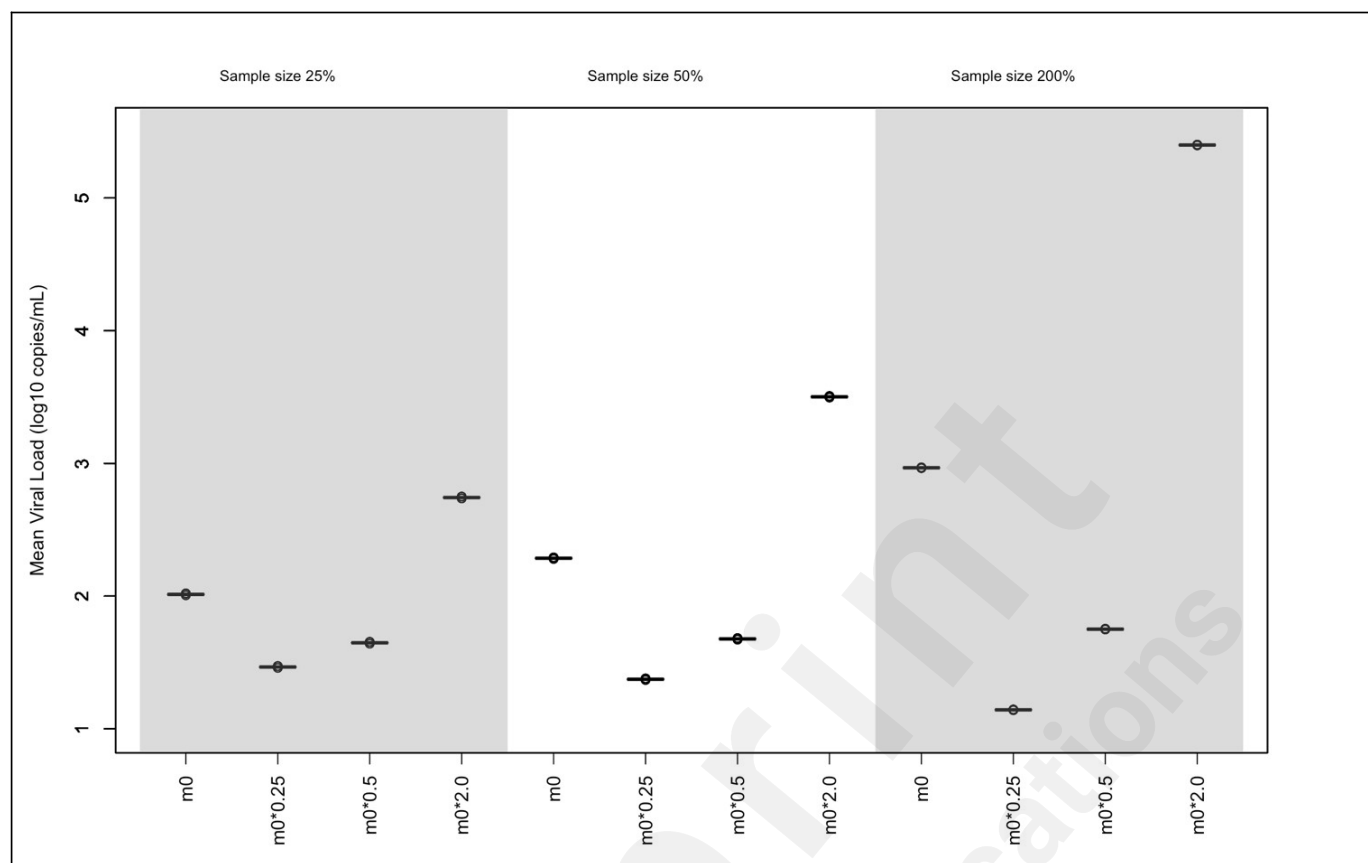
Discussion

Using a synthetic population, we observed that recovery of population VL from a single center's monitored VL was hampered when a historic measurement or informed guess at the prior population VL was unavailable. In other words, good VL data are preferred to methodological adjustments of incomplete data.

Community VL, calculated from individuals who have been diagnosed with HIV, has been used to generalize risk of HIV transmission and evaluate retention in care and viral suppression [6]. However, as mentioned in the introduction, this measure has several shortcomings. First, it will almost always result in underestimated VL as individuals who are unaware of their HIV status will likely have higher viral loads. Relatedly, timing of the individual VL measure may also impact the community estimate as VL will fluctuate over time (e.g., acute vs. chronic infection). Second, there may be issues with defining the specific geographic area of the community and whether this population is "closed." While closed communities would allow for a more accurate community VL

measurement, the applicability and feasibility are hindered in the real world by population migration. Third, sampling bias may be present when there is a high prevalence of undiagnosed PLWH [6]. In these situations, the use of the population VL may be more appropriate for reflecting transmission potential should we be able to impute data for those undiagnosed or not retained in care. To address these limitations, alternate metrics have been proposed, such as the prevalence of viraemia based on viral suppression [19]. As such, researchers have adopted alternative community level VL measures that reflect the prevalence of HIV in the community as well as distinguishing between those who are virally suppressed and those who have a high VL and are more likely to contribute to community spread [6, 19]. The methods we have demonstrated can readily be adapted to other HIV measures where a weighted mean may be desired, such as CD4 cell counts among PLWH for a given jurisdiction. Regardless of the metric used, there is still risk of ecologic fallacy at the aggregate level wherein a higher population VL may not correspond to higher individual transmission risk when prophylaxis is common.

Others have acknowledged the important challenge on the use of EHR data for population inference when healthcare seeking behavior and access to care impacts representativeness. EHR-based studies are susceptible to issues of confounding, information bias, and selection bias [9]. Bower et al. demonstrated how selection into an EHR is not random and recommend techniques such as sampling (post-stratification) weights and propensity scoring and inverse probability weighting (IPW) to adjust estimates, in their case, of cardiovascular disease risk [20]. Flood et al. used EHR data to estimate childhood obesity and found that the application of sampling weights to their data allowed estimates to be comparable to a nationally representative survey [21]. Goldstein et al. used IPW to adjust for presumed selection bias in single-center EHR-based study when exposure and outcome relate to catchment [22]. It is worth delineating how these two complementary strategies – sampling/post-stratification weighting versus propensity scoring/IPW – differ in EHR research.

The IPW approach requires specification of a probability model (i.e., the propensity score) for

selection into the EHR from the source population, conditioned on measured characteristics related to this process. However, this demands the EHR capture relevant details on the catchment process, or those data can be readily linked, and EHRs are well-known to lack data on epidemiological determinants [23]. On the other hand, using a sampling weight assumes we have access to the denominators from which the EHR data are sampled. One such source of data we have used are census estimates, which can be stratified by factors relating to catchment and tuned to the local environment. The challenge with this approach is that in practice we may not know all the characteristics defining catchment process, the census might not capture those characteristics, or there may be ambiguous geography. Indeed, catchment is a multifactorial and sometimes nebulous process related to healthcare availability, accessibility, affordability, accommodation, and acceptability [24]. One potential way to gain insight into catchment is to compare EHR data to census data to see which characteristics are over- or under-represented for a given geographic area defined by the clinic. If the census lacks data on catchment-relevant factors but the EHR captures these details (e.g., sexual orientation), this may favor the IPW approach.

Another important limitation of our approach was our construction of the sampling weights. We assumed a simple random sample within each catchment stratum to calculate the sampling weights. In our synthetic population, this was known with certainty, although we blinded ourselves to this oracle view by not retaining the selection probabilities during the data generation process but rather relying on our catchment model. However, as exemplified in our biased weighted analysis and the real-world clinic dataset, the catchment stratum may be uncertain, and in our case presumably underestimated population VL. Many extensions exist to improve weighting approaches, such as raking, which we did not evaluate herein [25]. We also observed a decrease in precision – widening of confidence intervals – when comparing the weighted versus unweighted results. This has been termed the bias-variance tradeoff, where improved accuracy may be accompanied by worsened precision [26].

A particular strength to our approach is the straightforward implementation and Bayesian adjustment that can be carried out with minimal programming ability. The included source code [18] can serve as a starting point. More complicated cluster survey designs may also benefit from Bayesian methods [27, 28]. Bayesian analysis requires careful deliberation over which priors may be most appropriate. Informative priors are useful and straightforward but obtaining unbiased estimates of VL can be prohibitively expensive for some jurisdictions, and measures obtained in one jurisdiction may not be exchangeable with another. Indeed, we observed that our real-world application was sensitive to the choice of prior. Nonetheless, even a small unbiased survey can dramatically reduce RMSE and thus may be justified. This would have to be done only once to seed Bayesian prospective surveillance of population VL. These methods can be adapted to other aggregated measures of disease prevalence, for both research and practice purposes, especially if an historic prior estimate is available.

Health departments have expressed interest in using EHR data for many community health measures than can help inform resource allocation and public health decision-making in different contexts. Comer et al. identified 23 of these [2]; hepatitis B and C infection was a high priority measure and one in which previous surveys such as the National Health and Nutrition Examination Survey [29, 30] can serve as an informative prior. If, for example, a focal outbreak of hepatitis C is detected from an EHR this could suggested targeted treatment and prevention efforts to cure infection and reduce future transmission.

In short, we observed that methodological adjustments were ineffective to recover the true population VL in our data without prior knowledge of what this value may be. Further validation using real-world EHR data from diverse clinical settings is needed to confirm this finding. Should such prior data be available, then it may be possible to infer population characteristics from a biased clinic sample in the EHR. Moving forward, we encourage those with access to population-based surveys of community health metrics – especially at sub-national levels – to continue to disseminate

these data to enable epidemiologic methods such as ours.

Acknowledgements

Conflicts of Interest

NDG consults for ChristianaCare, unrelated to the present work. All other authors report no financial conflicts of interest.

Abbreviations

EHR: electronic health record

GM: geometric mean

IPW: inverse probability weighting

PLWH: people living with HIV

RMSE: root mean squared error

VL: viral load

References

1. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. J Med Syst. 2018 Sep 29;42(11):214.

2. Comer KF, Gibson PJ, Zou J, Rosenman M, Dixon BE. Electronic Health Record (EHR)-Based Community Health Measures: An Exploratory Assessment of Perceived Usefulness by Local Health Departments. BMC Public Health. 2018 May 22;18(1):647.

3. U.S. Department of Health and Human Services. HIV National Strategic Plan for the United States: A Roadmap to End the Epidemic 2021–2025. 2021. HIV National Strategic Plan: 2021–2025.

4. Eisinger RW, Dieffenbach CW, Fauci AS. HIV Viral Load and Transmissibility of HIV Infection: Undetectable Equals Untransmittable. JAMA. 2019 Feb 5;321(5):451-452.

5. National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. Division of HIV/AIDS Prevention. Guidance on community viral load: a family of measures, definitions, and method for calculation. https://stacks.cdc.gov/view/cdc/28147. Published August 2011. Accessed May 5, 2023.

6. Miller WC, Powers KA, Smith MK, Cohen MS. Community viral load as a measure for assessment of HIV treatment as prevention. Lancet Infect Dis. 2013 May;13(5):459-64.

7. Chakraborty H, Weissman S, Duffus WA, Hossain A, Varma Samantapudi A, Iyer M, et al. HIV community viral load trends in South Carolina. Int J STD AIDS. 2017 Mar;28(3):265-

276.

8. Arey AL, Cassidy-Stewart H, Kurowski PL, Hitt JC, Flynn CP. Evaluating HIV Surveillance Completeness Along the Continuum of Care: Supplementing Surveillance With Health Center Data to Increase HIV Data to Care Efficiency. J Acquir Immune Defic Syndr. 2019 Sep 1;82 Suppl 1:S26-S32.

9. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol. 2021 Oct 27;21(1):234.

10. Centers for Disease Control and Prevention. HIV Surveillance Report, 2020; vol. 33. https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published May 2022. Accessed May 5, 2023.

11. Das M, Chu PL, Santos GM, Scheer S, Vittinghoff E, McFarland W, et al. Decreases in community viral load are accompanied by reductions in new HIV infections in San Francisco. PLoS One. 2010 Jun 10;5(6):e11068.

12. Schwarcz S, Hsu L, Dilley JW, Loeb L, Nelson K, Boyd S. Late diagnosis of HIV infection: trends, prevalence, and characteristics of persons whose HIV diagnosis occurred within 12 months of developing AIDS. J Acquir Immune Defic Syndr. 2006 Dec 1;43(4):491-4.

13. San Francisco Department of Public Health. HIV Epidemiology Annual Report 2021. August 2022. Available at: https://sfdph.org/dph/files/reports/RptsHIVAIDS/AnnualReport2021-Red.pdf. Accessed May 5, 2023.

14. Anderson AN, Higgins CM, Haardörfer R, Holstad MM, Nguyen MLT, Waldrop-Valverde D. Disparities in Retention in Care Among Adults Living with HIV/AIDS: A Systematic Review. AIDS Behav. 2020 Apr;24(4):985-997.

15. Chen C, Wakefield J, Lumely T. The use of sampling weights in Bayesian hierarchical models for small area estimation. Spat Spatiotemporal Epidemiol. 2014 Oct;11:33-43.

16. U.S. Census Bureau. 2021 American Community Survey 5-year Estimates, Table DP05.

Available at: https://data.census.gov/table?q=Delaware&tid=ACSDP1Y2021.DP05. Accessed May 5, 2023.

17. Abdallah S, Berl T, Dowling J, Pruitt G, Vella B, Vella N, et al. Delaware Integrated HIV Prevention and Care Plan: CY2022-2026. 2022. Available at: https://www.dhss.delaware.gov/ dph/dpc/files/comphivplan.pdf. Accessed May 5, 2023.

18. GitHub. goldsteinepi/population_vl. Available at: https://doi.org/10.5281/zenodo.11061525.

19. Herbeck J, Tanser F. Community viral load as an index of HIV transmission potential. Lancet HIV. 2016 Apr;3(4):e152-4.

20. Bower JK, Patel S, Rudy JE, Felix AS. Addressing Bias in Electronic Health Record-Based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise. Curr Epidemiol Rep. 2017 Dec;4(4):346-352.

21. Flood TL, Zhao YQ, Tomayko EJ, Tandias A, Carrel AL, Hanrahan LP. Electronic health records and community health surveillance of childhood obesity. Am J Prev Med. 2015 Feb;48(2):234-240.

22. Goldstein ND, Kahal D, Testa K, Burstyn I. Inverse probability weighting for selection bias in a Delaware community health center electronic medical record study of community deprivation and hepatitis C prevalence. Ann Epidemiol. 2021 Aug;60:1-7.

23. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. J Am Med Inform Assoc. 2020 Nov 1;27(11):1764-1773.

24. Penchansky R, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. Med Care. 1981 Feb;19(2):127-40.

25. Pew Research Center. January, 2018. "For Weighting Online Opt-In Samples, What Matters Most?" Available at: https://www.pewresearch.org/methods/2018/01/26/for-weighting-online- opt-in-samples-what-matters-most/. Accessed June 21, 2023.
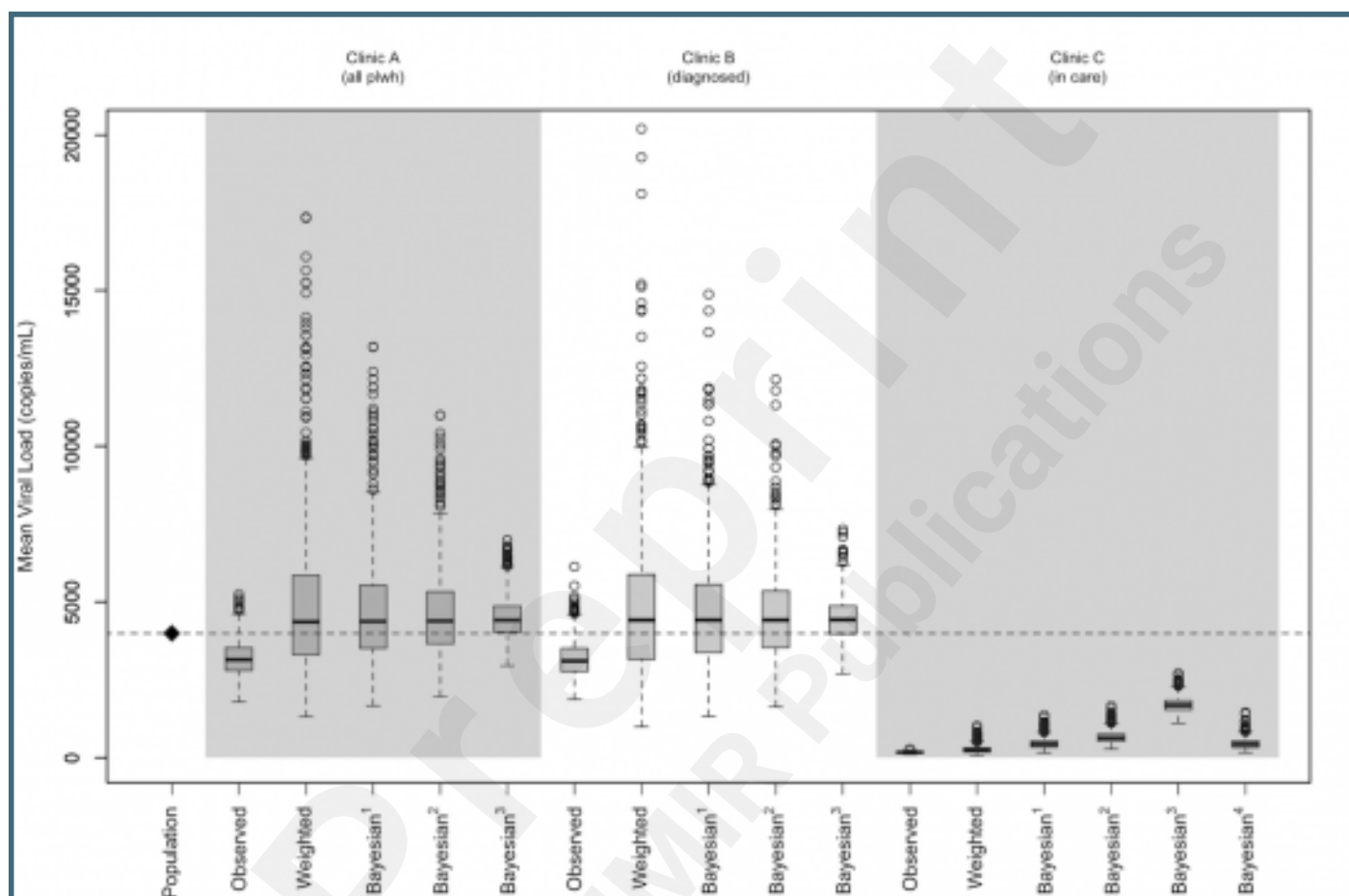
26. Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Chapman and Hall/CRC; 2003 Sep 25.

27. Makela S, Si Y, Gelman A. Bayesian inference under cluster sampling with probability proportional to size. Stat Med. 2018 Nov 20;37(26):3849-3868.

28. Zhou H, Elliott MR, Raghunathan TE. Multiple Imputation in Two-Stage Cluster Samples Using The Weighted Finite Population Bayesian Bootstrap. J Surv Stat Methodol. 2016 Jun 1;4(2):139-170.

29. Hofmeister MG, Rosenthal EM, Barker LK, Rosenberg ES, Barranco MA, Hall EW, Edlin BR, Mermin J, Ward JW, Ryerson AB. Estimating Prevalence of Hepatitis C Virus Infection in the United States, 2013-2016. Hepatology. 2019 Mar;69(3):1020-1031.

30. Roberts H, Kruszon-Moran D, Ly KN, Hughes E, Iqbal K, Jiles RB, Holmberg SD. Prevalence of chronic hepatitis B virus (HBV) infection in U.S. households: National Health and Nutrition Examination Survey (NHANES), 1988-2012. Hepatology. 2016 Feb;63(2):388-97.
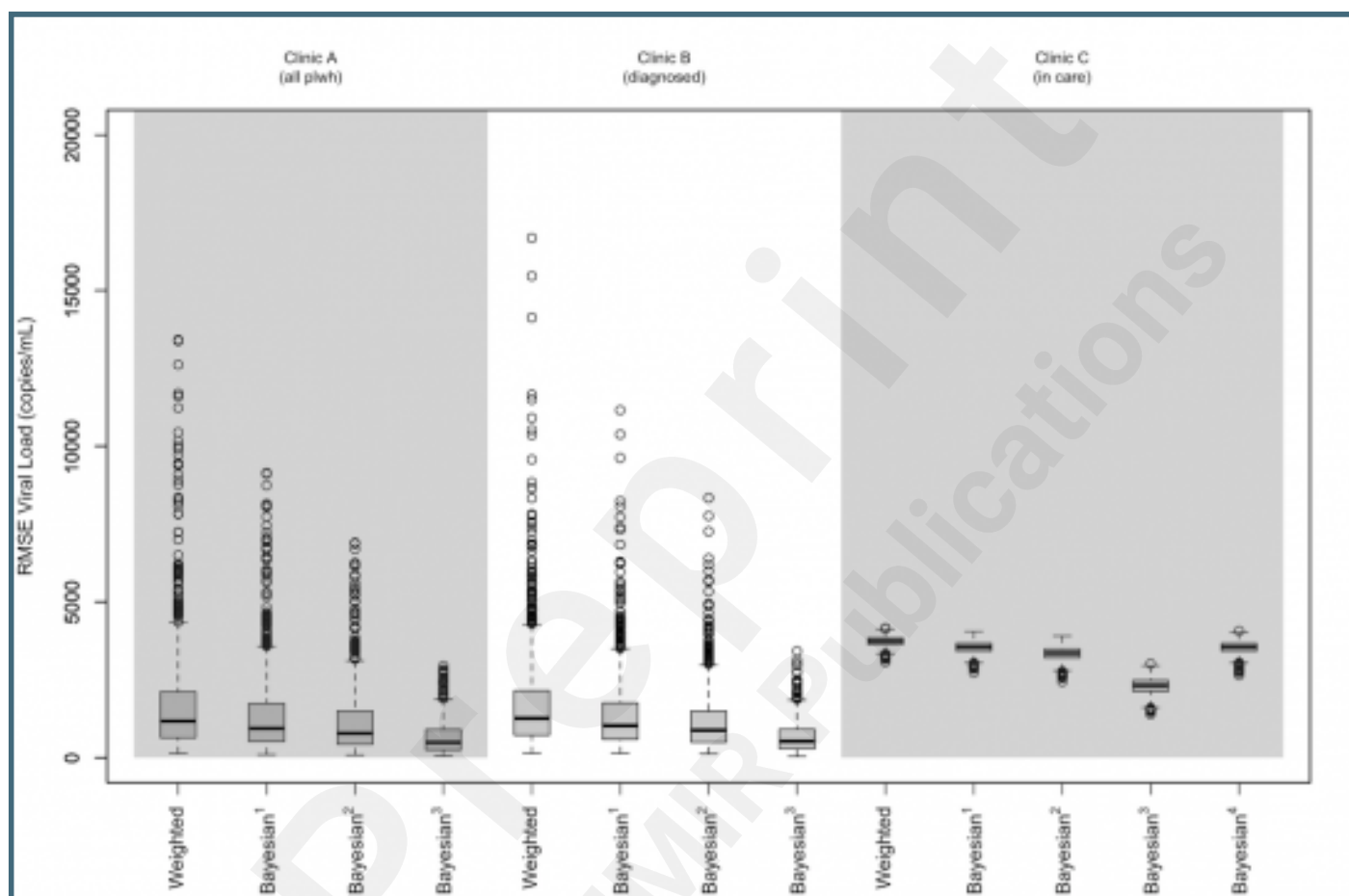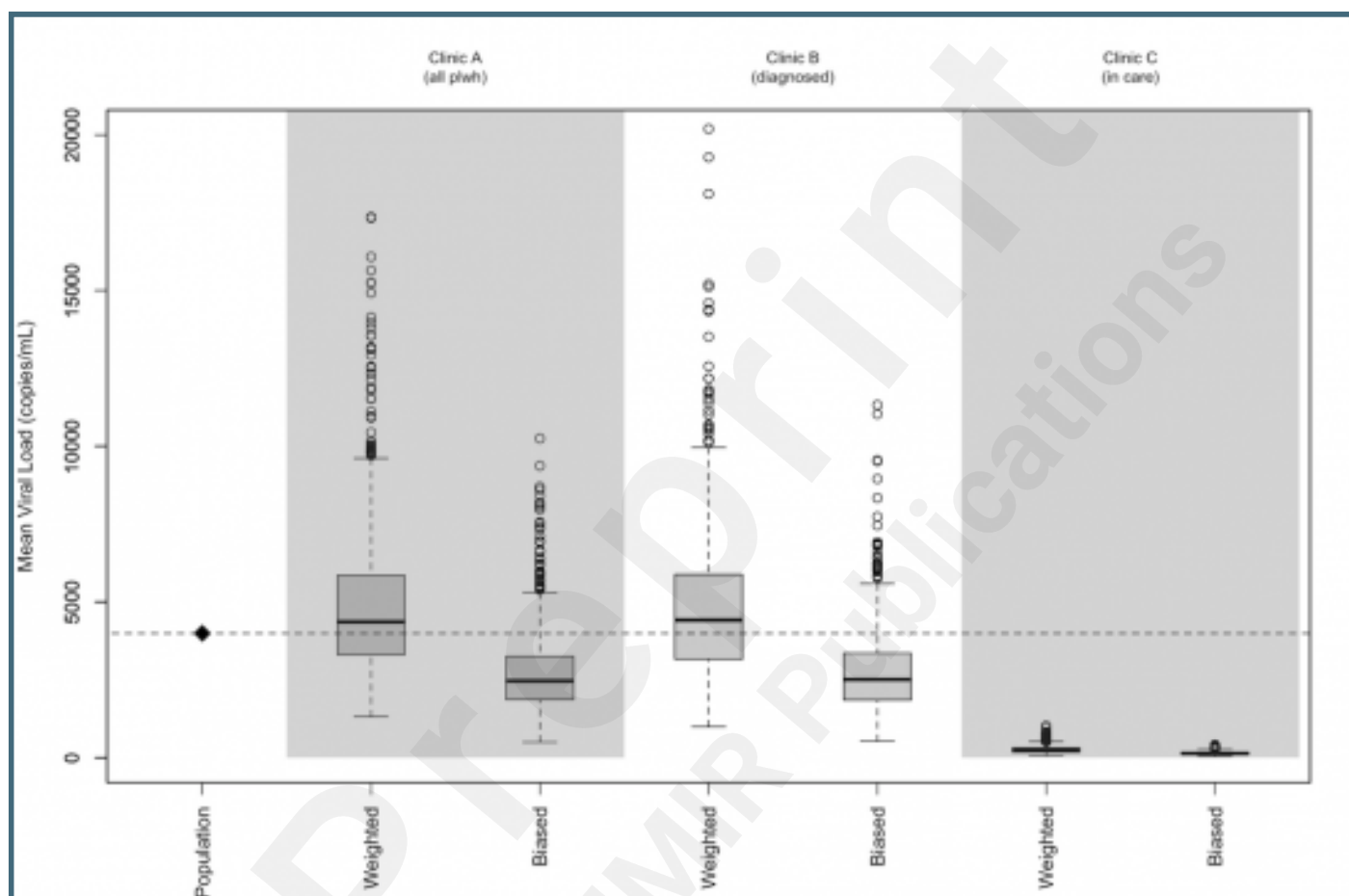
Supplementary Files

Figures

Comparison of the distribution of the geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. 1 Prior sample size of 25% of the weighted clinic sample size. 2 Prior sample size of 50% of the weighted clinic sample size. 3 Prior sample size of 200% of the weighted clinic sample size. 4 Prior mean informed from weighted clinic B estimates.

Comparison of the root mean squared error (RMSE) of the geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. 1 Prior sample size of 25% of the weighted clinic sample size. 2 Prior sample size of 50% of the weighted clinic sample size. 3 Prior sample size of 200% of the weighted clinic sample size. 4 Prior mean informed from weighted clinic B estimates.
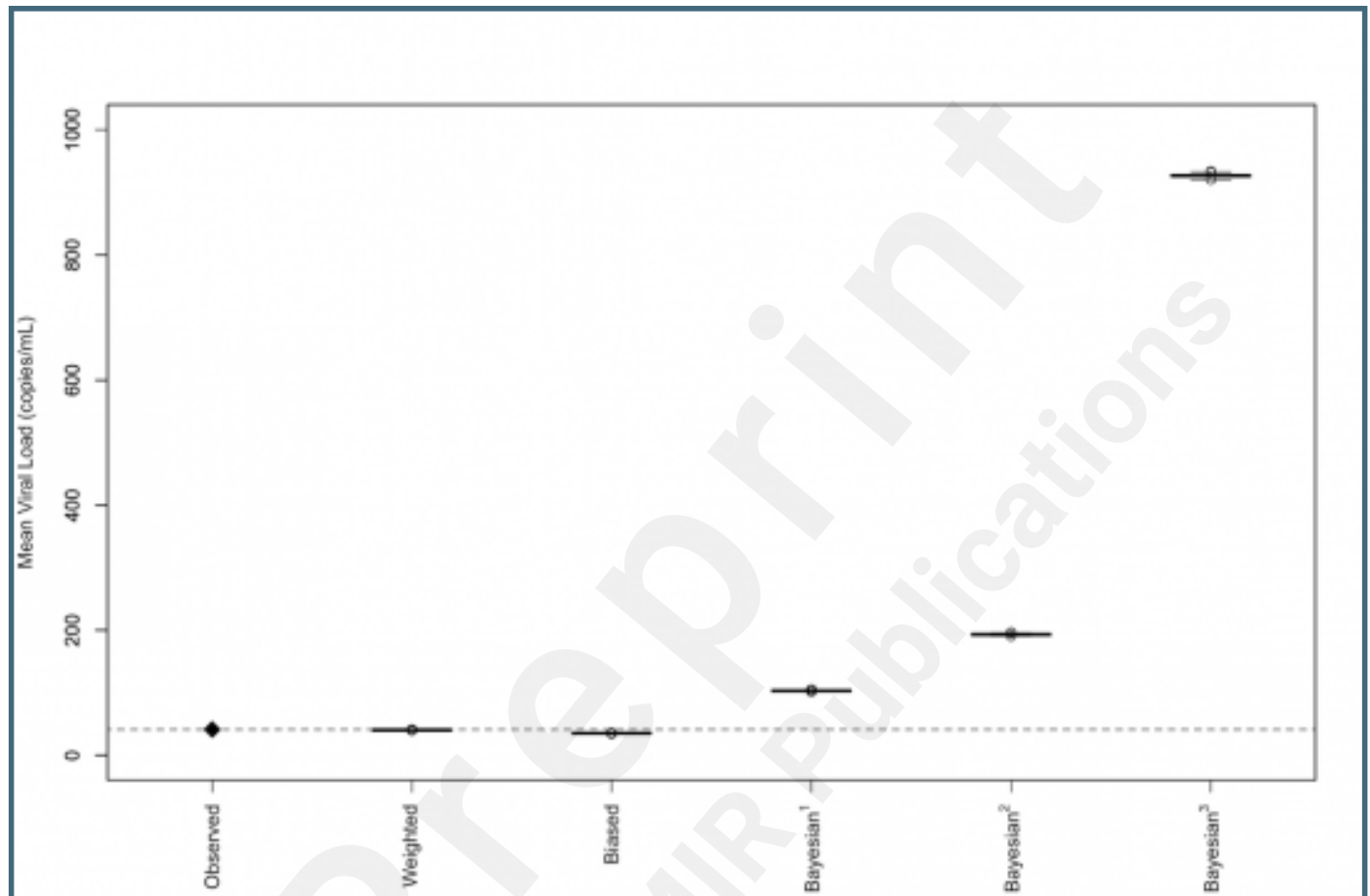
Comparison of weight misspecification in the weighted geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic.

Inferred population geometric mean HIV viral load (VL) for Delaware based on active patients retained in care at the Holloway Community Program at ChristianaCare (Wilmington, DE). Results represent 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic. 1 Prior sample size of 25% of the weighted clinic sample size. 2 Prior sample size of 50% of the weighted clinic sample size. 3 Prior sample size of 200% of the weighted clinic sample size.

Multimedia Appendixes

Comparison of the distribution of the logarithmic geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. 1 Prior sample size of 25% of the weighted clinic sample size. 2 Prior sample size of 50% of the weighted clinic sample size. 3 Prior sample size of 200% of the weighted clinic sample size. 4 Prior mean informed from weighted clinic B estimates.
URL: http://asset.jmir.pub/assets/906465eb8526f12abe4dab11d922e7e5.png

Comparison of the root mean squared error (RMSE) of the logarithmic geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. 1 Prior sample size of 25% of the weighted clinic sample size. 2 Prior sample size of 50% of the weighted clinic sample size. 3 Prior sample size of 200% of the weighted clinic sample size. 4 Prior mean informed from weighted clinic B estimates.
URL: http://asset.jmir.pub/assets/3a0238bf62f0b99f0bb3c9426f394184.png

Comparison of weight misspecification in the weighted logarithmic geometric mean HIV viral load (VL) for three clinic electronic health records (n=250 per clinic) sampled from a synthetic source population of people living with HIV (n=10,000). Clinic A was sampled directly from the source population, whereas clinics B and C were sampled from a subset of the source population based on diagnosed HIV (clinic B) or retention in care (clinic C). All synthetic clinics oversampled by male sex, White race, and 45 years of age or older. Results represent 1,000 hypothetical clinics each with 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic.
URL: http://asset.jmir.pub/assets/aaece205d5451a0417dbe6a2643204ed.png

Inferred population geometric mean HIV viral load (VL) for Delaware based on active patients retained in care at the Holloway Community Program at ChristianaCare (Wilmington, DE). Results represent 1,000 sampling weight adjustments. The biased weight systematically down-weighted higher VL individuals when they should have been up-weighted, as might occur based on an inaccurate catchment model where individuals with higher VLs were less likely to be sampled in the clinic. 1 Prior sample size of 25% of the weighted clinic sample size. 2 Prior sample size of 50% of the weighted clinic sample size. 3 Prior sample size of 200% of the weighted clinic sample size.
URL: http://asset.jmir.pub/assets/06260ded8a9dfe84e20942e40eb91d55.png

Sensitivity analysis of inferred population geometric mean HIV viral load (VL) for Delaware based on active patients retained in care at the Holloway Community Program at ChristianaCare (Wilmington, DE). Results represent 1,000 Bayesian sampling weight adjustments. Sensitivity analysis compared three alternate specifications of the prior mean for VL: 25%, 50%, and 200% of the original specification (m0). Prior sample size was varied three ways: 25%, 50%, and 200% of the weighted clinic sample size.
URL: http://asset.jmir.pub/assets/a646cac5ec40c523e6884c7964f644d6.png