BMJ Public Health

Addressing selection biases within electronic health record data for estimation of diabetes prevalence among New York City young adults: a cross-sectional study

Sarah Conderino ,[1] Lorna E Thorpe,[1] Jasmin Divers,[1,2] Sandra S Albrecht,[3] Shannon M Farley,[4] David C Lee,[1] Rebecca Anthopolos[1]

To cite: Conderino S, Thorpe LE, Divers J, *et al*. Addressing selection biases within electronic health record data for estimation of diabetes prevalence among New York City young adults: a cross-sectional study. *BMJ Public Health* 2024;**2**:e001666. doi:10.1136/bmjph-2024-001666

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjph-2024-001666).

Received 27 June 2024
Accepted 8 October 2024

[1]Department of Population Health, NYU Grossman School of Medicine, New York, New York, USA
[2]Department of Foundations of Medicine, NYU Long Island School of Medicine, Mineola, New York, USA
[3]Department of Epidemiology, Mailman School of Public Health at Columbia University, New York, New York, USA
[4]ICAP at Columbia University, New York, New York, USA

Correspondence to
Dr Sarah Conderino; sarah.conderino@nyulangone.org

ABSTRACT

Introduction There is growing interest in using electronic health records (EHRs) for chronic disease surveillance. However, these data are convenience samples of in-care individuals, which are not representative of target populations for public health surveillance, generally defined, for the relevant period, as resident populations within city, state or other jurisdictions. We focus on using EHR data for the estimation of diabetes prevalence among young adults in New York City, as the rising diabetes burden in younger ages calls for better surveillance capacity.

Methods This article applies common non-probability sampling methods, including raking, post-stratification and multilevel regression with post-stratification, to real and simulated data for the cross-sectional estimation of diabetes prevalence among those aged 18–44 years. Within real data analyses, we externally validate city-level and neighbourhood-level EHR-based estimates to gold-standard estimates from a local health survey. Within data simulations, we probe the extent to which residual biases remain when selection into the EHR sample is non-ignorable.

Results Within the real data analyses, these methods reduced the impact of selection biases in the citywide prevalence estimate compared with the gold standard. Residual biases remained at the neighbourhood-level, where prevalence tended to be overestimated, especially in neighbourhoods where a higher proportion of residents were captured in the sample. Simulation results demonstrated these methods may be sufficient, except when selection into the EHR is non-ignorable, depending on unmeasured factors or on diabetes status.

Conclusions While EHRs offer the potential to innovate on chronic disease surveillance, care is needed when estimating prevalence for small geographies or when selection is non-ignorable.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Electronic health records (EHRs) are a compelling data source for public health research but are prone to selection biases.

WHAT THIS STUDY ADDS

⇒ We use bias adjustment methods to estimate diabetes prevalence among children and young adults in New York City and demonstrate how these methods can be used to produce EHR-based prevalence estimates that are statistically equivalent to survey estimates.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ EHRs can inform chronic disease surveillance efforts. However, residual biases may exist if selection into the EHR depends on unmeasured factors.

INTRODUCTION

Increasingly, public health researchers and practitioners have explored how electronic health records (EHRs) can be leveraged for valid and reliable public health surveillance purposes.[1 2] While EHRs offer a compelling opportunity for surveillance, patient populations may be non-representative of the general population with respect to demographic characteristics.[3] From a health status perspective, patients represented within EHR data are typically sicker than the general population.[4] These differences introduce the potential for selection bias in EHR-based surveillance.[5]

Addressing selection bias in EHR-based surveillance is a formidable challenge. Contrary to a complex survey sample with known sampling weights to infer from the sample to the target, EHR data are non-probability samples wherein the process by which individuals select the sample is unknown. The statistical missing data lexicon has been adapted to characterise the selection

process in non-probability samples.[6] Selection completely at random describes scenarios whereby each individual has an equal probability of selection into the sample. Selection at random (SAR) describes scenarios whereby the probability of selection depends on observed characteristics of the individuals, but given those characteristics, is independent of unobserved outcomes from individuals absent from the sample.[7–9] Lastly, selection not at random (SNAR) refers to selection processes whereby the probability of selection is dependent on unobserved outcomes, even after adjusting for observed covariates.[7–9] For valid EHR-based surveillance, the selection mechanism for the EHR sample needs to be incorporated into the estimation approach.

Previous research for EHR-based surveillance has used various non-probability sampling methods to estimate population disease prevalence.[10–12] Based on SAR, these methods assume that after controlling for variables captured in the EHR sample and population, such as basic demographics, the selection process no longer depends on the unobserved disease status of individuals not represented in the EHR sample. However, the tendency for EHRs to over-represent sicker individuals increases the plausibility of SNAR, suggesting the assumptions behind SAR are unlikely to be correct. As the goal of surveillance is estimation in the general population, including those not in-care individuals, this type of SNAR scenario can contribute to the overestimation of disease prevalence and incidence. The extent to which EHR-derived surveillance estimates may be sensitive to SAR assumptions has received little attention in previous literature.[10]

As part of wider efforts to use EHR data to estimate diabetes prevalence among young adults,[13] a population that is experiencing rising diabetes burden,[14] we conducted a multi-step process to evaluate common bias adjustment methods. First, we conducted a data illustration using real data where we could evaluate validity against 'gold-standard' estimates. Second, we conducted simulations where we could generate various selection processes to explore hypothesised factors that could contribute to residual biases observed within the initial data illustration. The overarching goal of the paper was to compare these bias adjustment methods using real data and simulations to help inform the broader discussion on how to effectively use EHRs for population-level surveillance purposes.

METHODS

Data illustration

NYU Langone Health (NYU) is a large academic medical centre that serves patients throughout New York City (NYC). NYU includes three major hospitals, an extensive network of outpatient clinics and one of the nation's largest Federally Qualified Health Center networks. Longitudinal NYU EHR data were obtained for all NYC-resident patients aged 18–44 years with an inpatient or outpatient encounter from 2017 to 2019. The main analyses included all NYC residents since prevalence estimation to the full NYC jurisdiction is of greater public health relevance. As some researchers have attempted to limit EHR samples to health system service areas to reflect primary populations served by their facilities and to potentially reduce selection biases,[13] we conducted sensitivity analyses varying the resident inclusion criteria to restrict NYC neighbourhoods within different definitions of NYU service areas (online supplemental appendix).

Using EHR data through 2019, we defined patients with diabetes as those with ≥2 diagnoses for diabetes, 1 diagnosis and ≥2 elevated A1C laboratory studies ≥6.5% or at least 1 anti-diabetes prescription (excluding metformin/acarbose).[15] Demographic variables defined in the EHR sample included age group, sex, race/ethnicity, Medicaid insurance status and Public Use Microdata Areas (PUMAs), Census subgeographies containing ≥100 000 residents to proxy neighbourhood of residence (n=55). Race/ethnicity was imputed for those with unknown race/ethnicity (19%) using the Bayesian Improved Surname Geocoding (BISG) methods.[16] All patients with an unknown/other age or sex were excluded (<1%). To characterise the demographics of the target population, we defined equivalent demographic variables on the NYC subset of American Community Survey (ACS) 2019 5-year data obtained through IPUMS USA, a line-level sample of ACS data that is weighted to the general population.[17]

We estimated diabetes prevalence overall and by PUMAs according to four estimation methods: crude, raking, post-stratification and multilevel regression with post-stratification (MLRP). In the crude method, we calculated the proportion of patients within the EHR sample who were classified as having diabetes. In the raking method, we iteratively adjusted the EHR sample to match the marginal distribution of demographic covariates in the general population.[18] In the post-stratification method, we adjusted the EHR sample to match the joint distribution of demographic covariates in the general population.[18] In the MLRP method, we fit a multilevel logistic regression model to predict diabetes in the EHR sample, including fixed effects for binary demographics and random effects for all non-binary individual-level demographics.[19 20] Full details on model specification and sensitivity analyses of alternative specifications that include neighbourhood-level social determinant of health (SDOH) and health outcomes are found in online supplemental appendix I. Model predicted probabilities were applied to the post-stratification weights within the general population.

The proxy gold standard prevalence estimates for comparison were calculated using pooled 2015–2020 data from the pooled Community District (CD) version of the NYC Community Health Survey (CHS).[21 22] The NYC CHS is an annual, cross-sectional telephone survey of a stratified random sample of approximately 10 000 NYC adults.[23] The pooled version includes respondents who are assigned to a CD, an NYC geographical unit that approximates PUMAs.[24] We compared EHR-derived

crude and adjusted prevalence estimates to diabetes prevalence estimates from external surveillance systems using three measures: (1) the relative difference from the gold standard estimate (P_{EHR}–P_{CHS})/P_{CHS} × 100); (2) statistical equivalence to the gold standard estimate through the two one-sided test (TOST) using an alpha of 0.05 and equivalence bounds of 0.005; and (3) the Pearson correlation coefficient between the neighbourhood-level EHR and gold standard estimates.

Simulation study

Based on the results from the data illustration, simulations were run to probe the extent to which residual biases remain under two SNAR scenarios: (1) selection is dependent on an unmeasured factor, for which there is proxy/auxiliary information measured in the sample and general population (eg, socioeconomic status [SES]); and (2) selection is dependent on diabetes status in the population.

Simulated populations were composed of 500 000 individuals equally distributed across 50 neighbourhoods to approximate the number of NYC PUMAs. Diabetes status and selection into the EHR sample were simulated using observed associations obtained through real-world data (online supplemental appendix). Diabetes mellitus ('DM') and selection were defined using mixed effects regression models with probit link functions. The DM model included fixed effects for demographic variables and random effects to generate heterogeneity in diabetes prevalence across neighbourhoods. The selection model included fixed effects for demographics, neighbourhood distance from the healthcare facility and random effects to generate heterogeneity in selection for the interaction of sex and race/ethnicity. Baseline associations between all variables are displayed in figure 1. Overall, the simulated populations had a true mean diabetes prevalence of 3% and a mean probability of selection into the sample of 10%.

Simulation scenario 1 introduced a binary individual-level, unobserved variable '*U*' that was associated with DM (OR=2.0) and selection (OR=0.7), which was modelled after observed patterns with household poverty level.[23] An observed auxiliary variable '*W*' was defined based on a set association with *U*, which was modified at levels equivalent to 10%, 30%, 50%, 70% and 90% misclassification when using *W* as a proxy for *U*. For scenario 1, *U* was not included in the adjustment procedures but *W* was. Simulation scenario 2 introduced and modified an association between DM and selection (OR_{DM}) at OR levels of 0.33, 0.67, 1.0, 1.5 and 3.0, selecting an upper limit from the crude association between DM and having a personal doctor/provider within CHS data. For each scenario, 100 simulations were run.

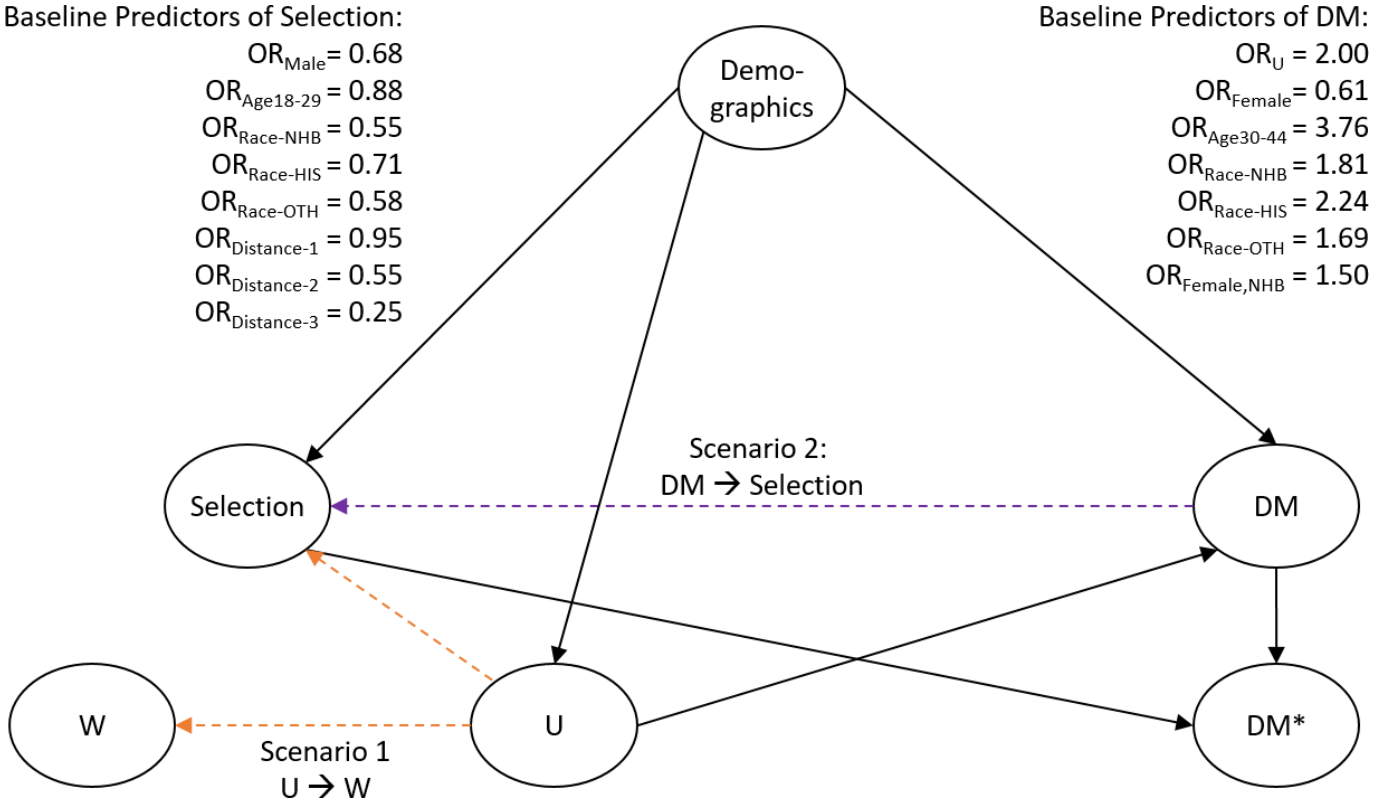

Figure 1 Simulation study directed acyclic graph with baseline OR associations. Observed diabetes within those selected into the EHR sample; scenario 1 (orange): modified the level of misclassification of the auxiliary variable W compared with the unobserved variable U at levels equivalent to 10%, 30%, 50%, 70% and 90% misclassification; scenario 2 (purple): modified the association between diabetes and selection at OR levels of 0.33, 0.67, 1.0, 1.5 and 3.0. DM, diabetes mellitus; EHR, electronic health record; HIS, Hispanic; NHB, non-Hispanic Black; OTH, Other race.

Table 1 Demographic profile of the NYU Langone EHR sample and NYC general population, young adults aged 18–44 years

	NYC general population*	NYU Langone EHR sample	Crude EHR-based diabetes prevalence
Sex			
Female	51.2%	62.2%	2.93%
Male	48.8%	37.8%	3.35%
Race			
Black	20.3%	12.7%	4.23%
Latino	29.6%	19.1%	4.44%
Other	18.1%	16.1%	2.88%
White	32.0%	52.1%	2.38%
Age			
18–29	43.6%	37.5%	1.88%
30–44	56.4%	62.5%	3.82%
Insurance			
Non-Medicaid	74.2%	77.8%	2.78%
Medicaid	25.8%	22.2%	4.18%

*Defined using American Community Survey 2019 5-year data obtained through IPUMS USA.[17]
EHR, electronic health record; NYC, New York City; NYU, NYU Langone Health.

Simulations produced the true diabetes prevalence within the general population, crude prevalence within the EHR sample and estimated prevalence adjusted to the general population using raking, post-stratification and MLRP. We assessed the performance of each adjustment method using: (1) relative bias, or the average per cent difference between the true diabetes prevalence in the full population and the estimated diabetes prevalence within the sample; and (2) coverage probability, or the percentage of simulations with a true diabetes prevalence falling within the 95% CI. All analyses were performed using R V.4.1.2.[25]

Patient and public involvement

No patient involved.

RESULTS

Data results

A total of 454 612 young adults were identified in the EHR sample. Compared with the NYC general population, the EHR sample had over-representation of white (1.6-fold) and female (1.2-fold) individuals, who had a lower crude prevalence of diabetes than other racial/ethnic or sex subgroups (table 1). The sample also had over-representation of those aged 30–44 years (1.1-fold), who had a greater crude prevalence of diabetes than those aged 18–29 years (3.82% vs 1.88%). Representation varied more substantially across the 55 neighbourhoods (figure 2A).

According to the gold standard survey, diabetes prevalence among young adults was 3.33% (95% CI: 3.02 to 3.67) (table 2). Within the EHR sample, 3.09% were classified as having diabetes (95% CI: 3.04 to 3.14), 0.92 times the gold standard (−7.88% relative difference) and not statistically equivalent through the TOST. Adjusted prevalence estimates using raking, post-stratification and MLRP (ranging from 3.54% to 3.55%) were approximately 1.06 times the gold standard and statistically equivalent at the equivalence bound of 0.005, though improvements in relative differences were small (5.75%–6.16%). Prevalence estimates by race, age group and sex are presented in online supplemental appendix table 1. Subgroup estimates were comparable across adjustment methods.

When comparing EHR-based and gold standard prevalence estimates at the PUMA neighbourhood-level, there was a moderate, statistically significant correlation (R=0.5, p<0.001) for all EHR-based methods; though, as with the overall adjusted estimates, neighbourhood-level EHR estimates were generally higher than the neighbourhood-level gold standard estimates (figure 2B). In addition, as the proportion of the general population captured within the EHR sample increased, the relative difference from the gold standard estimates increased (figure 2C).

Sensitivity analyses varying the residential inclusion criteria to NYU service areas found that demographic representativeness of the sample increased within service areas where a greater proportion of the general population was captured in the sample (online supplemental appendix table 2). When these samples were adjusted and externally validated to the general population within the overall equivalent service area, they produced estimates that were systematically higher and not statistically equivalent to the gold standard estimate for the service area (online supplemental appendix tables 3 and 4). Sensitivity analyses including neighbourhood-level SDOH and health outcomes in the MLRP model did not meaningfully affect the overall or neighbourhood-level prevalence estimates (online supplemental table 3).

Simulation results

In scenario 1, crude diabetes prevalence within the simulated EHR sample had an average relative bias of approximately −40% when the unobserved variable U was introduced into the selection process (figure 3A). Adjustment methods including W partially accounted for this bias, however substantial residual biases remained, with coverage below 70% for all adjusted estimates (online supplemental table 5). The level of residual biases depended on the strength of the association between the auxiliary and unobserved variables but not on the direction. For both 10% and 90% misclassification, average relative biases were approximately −10%, and for both 30% and 70% misclassification, average relative biases were approximately −20%.

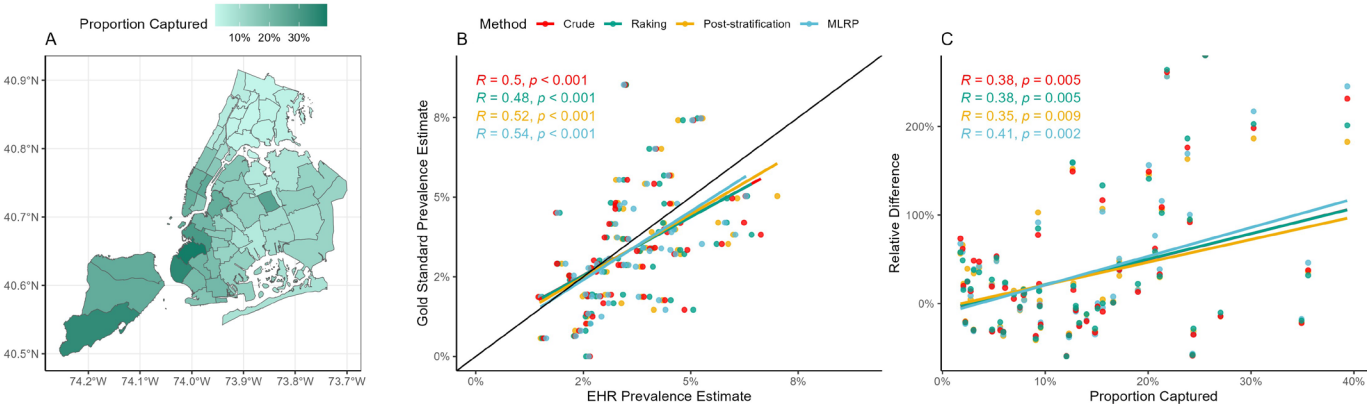

Figure 2 Characterisation of the NYU Langone patient sample and comparison of NYU EHR-based to gold standard diabetes prevalence estimates for young adults aged 18–44 years by New York City PUMA neighbourhood. (A) Proportion of general population captured within the EHR sample by NYC PUMA, calculated by dividing NYU Langone patient counts by the total NYC PUMA population estimates from the American Community Survey 2019 5-year data, obtained through IPUMS USA. (B) Comparison of NYU EHR-based to gold standard diabetes prevalence estimates. Each point represents a PUMA neighbourhood. EHR estimates are defined using NYU Langone Health 2019 data. The gold standard estimate is defined using NYC CHS 2015–2020 data. (C) Comparison of relative bias in NYU EHR-based prevalence estimates versus proportion of the general population captured within the EHR sample. Relative bias is calculated as the per cent change between the gold standard and EHR-based prevalence estimate for each NYC PUMA neighbourhood. CHS, Community Health Survey; EHR, electronic health record; MLRP, multilevel regression with post-stratification; NYC, New York City; NYU, NYU Langone Health; PUMA, Public Use Microdata Area.

In scenario 2, crude diabetes prevalence within the simulated EHR sample had an average relative bias ranging from −94% when those with diabetes had strongly decreased odds of selection (OR_{DM}=0.33) to +143% when those with diabetes had strongly increased odds of selection (OR_{DM}=3.0) (figure 3B). Adjustment methods did not have a meaningful impact on the residual biases when those with diabetes had decreased odds of selection (OR_{DM}=0.33 or OR_{DM}=0.67), with coverage at 0% for all methods. When those with diabetes had increased odds of selection (OR_{DM}=1.5 or OR_{DM}=3.0), adjustment methods increased the relative bias compared with crude estimates (figure 3B). Simulation results displayed similar patterns for neighbourhood-level estimates (online supplemental appendix figure 1).

Table 2 Diabetes prevalence among NYC young adults 18–44 years, estimated from the NYU Langone Health electronic health record versus NYC Community Health Survey (NYC CHS) gold standard

	Prevalence (%) (95% CI)	Relative difference from gold standard (NYC CHS)*
Gold standard		
NYC CHS	3.33 (3.02 to 3.67)	–
EHR-based		
Crude	3.09 (3.04 to 3.14)	−7.88%
Raking	3.55 (3.46 to 3.63)	6.02%†
Post-stratification	3.54 (3.43 to 3.64)	5.75%†
MLRP	3.55 (3.47 to 3.63)	6.16%†

*Per cent difference from the gold standard estimate, the New York City Community Health Survey.
†Reject the null hypothesis of the TOST, or equivalent to the gold standard within equivalence bounds of 0.005.
EHR, electronic health record; MLRP, multilevel regression with post-stratification; NYC, New York City; NYU, NYU Langone Health; TOST, two one-sided test.

DISCUSSION

In this paper, bias adjustment methods were applied to EHR data to explore whether valid diabetes prevalence estimates could be generated for young adults within NYC. Within the NYU sample, the crude prevalence was lower than the proxy gold standard estimate of diabetes prevalence for NYC young adults, which may have been driven by demographic differences. Compared with the target population, the EHR sample had a higher proportion of female and white individuals, which are groups known to have lower diabetes prevalence. All adjustment methods performed similarly and produced prevalence estimates that were statistically equivalent to the gold standard, although systematically higher. Within neighbourhood-level analyses, we observed that relative differences from gold standard estimates increased as a proportion of the general population captured in the sample increased. Further, larger relative differences were observed in sensitivity analyses that were restricted to NYU service areas. These findings were counterintuitive, as these samples were more representative of the target populations based on measured demographics; we would assume that representativeness on unmeasured factors would also increase.

Simulation analyses were then used to probe the potential for residual selection biases within EHR-derived

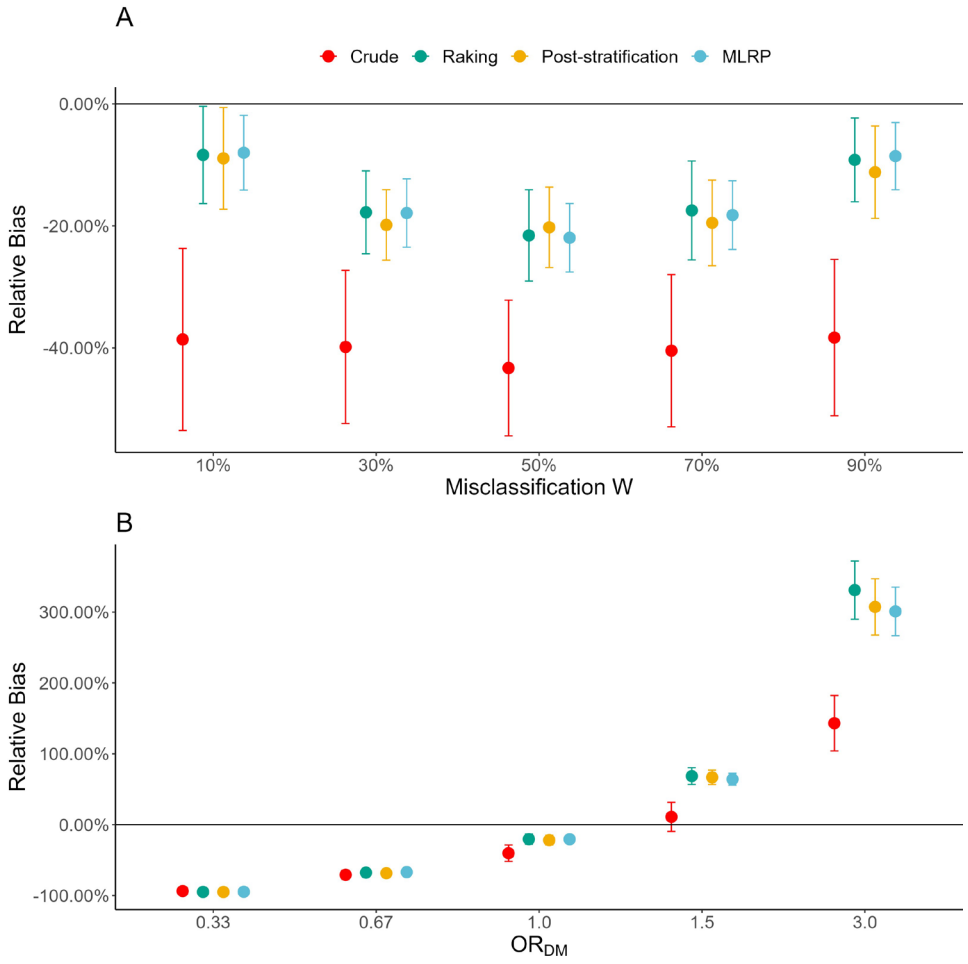

Figure 3 Mean relative bias in the EHR-based estimates versus true diabetes prevalence by simulation scenario. Error bars represent SD in mean relative bias across simulations. (A) Scenario 1 modified the level of misclassification of the auxiliary variable *W* compared with the unobserved variable *U*; (B) scenario 2 modified the association between diabetes and selection (OR_{DM}). EHR, electronic health record; MLRP, multilevel regression with post-stratification.

estimates. Scenario 1 introduced an unobserved predictor of diabetes for which there was an imperfect proxy variable. This scenario demonstrated that residual biases may still exist even when this auxiliary information is a strong proxy of unobserved predictors. Evidence supports that SDOH is associated with diabetes and healthcare utilisation.[3] However, these variables are notoriously difficult to measure using EHR data. Consistent with prior research, Medicaid status and neighbourhood-level SDOH were imperfect proxies that may not have fully accounted for selection biases by these factors in our data illustration.[26 27] Continued efforts to incorporate and use SDOH screening tools within EHRs may improve estimation through these methods.[28]

Scenario 2 introduced an association between diabetes and selection into the sample. Importantly, this scenario demonstrated that biases could be exacerbated through these methods when diabetes independently increased the odds of selection into the sample, which is plausible given individuals with chronic conditions are more likely to receive regular care than individuals who are healthy.[5 29 30] These selection biases could be further complicated by neighbourhood. For example, patients residing in neighbourhoods within close proximity, where the capture of the general population within the sample is high, may be more likely to use the health system for routine care, including diabetes management.[30] The observed positive relative differences and positive trends between relative differences and the proportion of the general population captured in the NYU sample could be partially attributed to such a mechanism. Including neighbourhood-level health outcomes in the MLRP models did not have a large impact on prevalence estimates, consistent with prior research using neighbourhood hospitalisation rates.[10] As proposed in the missing data literature, additional granular data on variables that are strongly correlated with diabetes (eg, obesity) within the general population could improve these methods.[31]

In this analysis, using common non-probability sampling methods to adjust for demographic non-representativeness of the EHR sample was effective in reducing the impact of selection biases in the overall estimate of diabetes prevalence among NYC young adults. However, based on the data illustration and simulation analyses, these methods, as implemented, may not always consistently produce valid estimates of diabetes

prevalence across jurisdictions or EHR sources. Observed positive relative differences compared with gold standard estimates support the hypothesised presence of an SNAR mechanism, where those with diabetes are more likely to be users of healthcare systems. This could contribute to an overestimation of diabetes prevalence, which could be exacerbated within certain neighbourhoods or other subgroups of interest. Of the methods tested in this work, MLRP has the greatest potential for addressing the more complex selection biases that are likely present within EHR data through the inclusion of auxiliary information within the predictive model. This potential could be realised by using population-representative clinical data sources (eg, all-payer claims databases) to incorporate neighbourhood-level healthcare utilisation patterns or health outcomes at more granular geographic scales.

This study has a number of strengths. The data illustration was conducted in NYC, a diverse, urban centre that includes several academic medical centres and a large system of 11 public hospitals; thus, the likelihood of any private health system being representative of the general population is low. NYC is also home to granular, high-quality gold standard data sources from external surveillance systems, allowing for validation of neighbourhood-level prevalence estimates. Comparison to a gold standard alone cannot facilitate understanding the conditions under which different methods will provide valid estimation. Our use of data simulations fills this gap by testing these methods under controlled conditions, which may inform the transportability of these methods to other populations or health outcomes.

However, there are several limitations to this analysis. NYC CHS gold standard estimates are based on self-reported diabetes status, which may under-report undiagnosed individuals not in-care. Further, variation in care-seeking patterns by demographic factors[3] could result in differential misclassification in self-reported diabetes status. The NYC CHS data were also pooled from 2015 to 2020 to produce reliable neighbourhood-level prevalence estimates within this age group. As diabetes prevalence has increased over time, this pooling could also contribute to lower prevalence within the gold standard. Additionally, while the computable phenotype for diabetes status was based on prior literature, it has not been validated within NYU. Misclassification of diabetes status may depend on healthcare utilisation patterns, which could contribute to positive relative differences observed within the data illustration.[5] Sensitivity analyses incorporating neighbourhood-level SDOH or health outcomes did not have a large impact on the results, which may have been driven by the use of PUMAs.[32] Neighbourhood-level factors defined using smaller geographical areas have been shown to improve the estimation of diabetes prevalence through MLRP methods.[10] While assumptions underlying the data generation process in the simulations were based on real-world data, these likely represent a simplification of true selection processes. Finally, race/ethnicity was missing on a large proportion of the population, and we relied on BISG imputation, which could have resulted in misclassification. BISG is also not feasible when using pseudonymised EHR data. EHRs offer a rich source of clinical information that can inform public health surveillance, yet selection biases inherent in these data can limit their utility, especially in generating small-area estimates. Statistical methods like MLRP can help account for these biases but depend on the ability to measure and adequately account for factors that affect selection into the EHR, which is likely to vary across jurisdictions and EHR data sources. Further, an understanding of underlying selection mechanisms is critical, as these methods have the potential to exacerbate biases. Future analyses should examine these issues for a variety of chronic diseases or locations, as selection biases likely differ across diseases, populations or EHR data sources.

Contributors SC contributed to study design, data acquisition, data analysis, data curation, drafting and editing of the manuscript. RA contributed to study design, data analysis and editing of the manuscript. JD, LET, SA, SMF and DCL contributed to study design and editing of the manuscript. All authors read and approved the final manuscript. SC is the guarantor.

Funding This work was supported by the Centers for Disease Control and Prevention (grant number 1U18DP006510).

Map disclaimer The depiction of boundaries on this map does not imply the expression of any opinion whatsoever on the part of BMJ (or any member of its group) concerning the legal status of any country, territory, jurisdiction or area or of its authorities. This map is provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the NYU Winthrop Hospital Institutional Review Board (i20-01338) and Columbia University Institutional Review Board (AAAU5390).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data are unavailable due to protected health information.

ORCID iD

Sarah Conderino http://orcid.org/0000-0003-1762-9262

REFERENCES

1 Perlman SE. Use and Visualization of Electronic Health Record Data to Advance Public Health. *Am J Public Health* 2021;111:180–2.

2 Kruse CS, Stein A, Thomas H, *et al*. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. *J Med Syst* 2018;42:1–16.
3 Queenan JA, Williamson T, Khan S, *et al*. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. *CMAJ Open* 2016;4:E28–32.
4 Romo ML, Chan PY, Lurie-Moroni E, *et al*. Characterizing Adults Receiving Primary Medical Care in New York City: Implications for Using Electronic Health Records for Chronic Disease Surveillance. *Prev Chronic Dis* 2016;13:E56.
5 Bower JK, Patel S, Rudy JE, *et al*. Addressing Bias in Electronic Health Record-Based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise. *Curr Epidemiol Rep* 2017;4:346–52.
6 Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–92.
7 Little RJ, Rubin DB. *Statistical Analysis with Missing Data*, 793. John Wiley & Sons, 2019.
8 Nandram B, Choi JW. Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data. *Surv Methodol* 2005;31:73–84. Available: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005001/article/8089-eng.pdf?st=uO3oK_PI
9 Little RJA, West BT, Boonstra PS, *et al*. Measures of the Degree of Departure from Ignorable Sample Selection. *J Surv Stat Methodol* 2020;8:932–64.
10 Chen T, Li W, Zambarano B, *et al*. Small-area estimation for public health surveillance using electronic health record data: reducing the impact of underrepresentation. *BMC Public Health* 2022;22:1515.
11 Thorpe LE, McVeigh KH, Perlman S, *et al*. Monitoring Prevalence, Treatment, and Control of Metabolic Conditions in New York City Adults Using 2013 Primary Care Electronic Health Records: A Surveillance Validation Study. *EGEMS (Wash DC)* 2016;4:1266.
12 Flood TL, Zhao YQ, Tomayko EJ, *et al*. Electronic health records and community health surveillance of childhood obesity. *Am J Prev Med* 2015;48:234–40.
13 Hirsch AG, Conderino S, Crume TL. Using electronic health records to enhance surveillance of diabetes in children, adolescents and young adults: a study protocol for the DiCAYA Network. *BMJ Open* 2024;14:e073791.
14 Wang L, Li X, Wang Z, *et al*. Trends in Prevalence of Diabetes and Control of Risk Factors in Diabetes Among US Adults, 1999-2018. *JAMA* 2021;326:704.
15 Avramovic S, Alemi F, Kanchi R, *et al*. US veterans administration diabetes risk (VADR) national cohort: cohort profile. *BMJ Open* 2020;10.
16 Imai K, Khanna K. Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Polit anal* 2016;24:263–72.
17 Ruggles S, Flood S, Goeken R, *et al*. IPUMS USA. In: *IPUMS*. Minneapolis, MN, 2022.
18 Lumley T. Analysis of Complex Survey Samples. *J Stat Soft* 2004;9:1–19.
19 Gelman A, Little TC. Poststratification into many categories using hierarchical logistic regression. 1997.
20 Bates D, Mächler M, Bolker B, *et al*. Fitting Linear Mixed-Effects Models Using lme4. *arXiv* 2014.
21 Hsia J, Zhao G, Town M, *et al*. Comparisons of Estimates From the Behavioral Risk Factor Surveillance System and Other National Health Surveys, 2011-2016. *Am J Prev Med* 2020;58:e181–90.
22 Bowlin SJ, Morrill BD, Nafziger AN, *et al*. Validity of cardiovascular disease risk factors assessed by telephone survey: the Behavioral Risk Factor Survey. *J Clin Epidemiol* 1993;46:561–71.
23 New York City Department of Health and Mental Hygiene. Community health survey restricted dataset. 2015-2020.
24 New York City Department of Planning. Community district profiles. Available: https://communityprofiles.planning.nyc.gov/about [Accessed 07 Aug 2023].
25 R: a language and environment for statistical computing [computer program]. version 4.1.2. Vienna, Austria R Foundation for Statistical Computing; 2010.
26 Casey JA, Pollak J, Glymour MM, *et al*. Measures of SES for Electronic Health Record-based Research. *Am J Prev Med* 2018;54:430–9.
27 Bhavsar NA, Gao A, Phelan M, *et al*. Value of Neighborhood Socioeconomic Status in Predicting Risk of Outcomes in Studies That Use Electronic Health Record Data. *JAMA Netw Open* 2018;1:e182716.
28 Cottrell EK, Dambrun K, Cowburn S, *et al*. Variation in Electronic Health Record Documentation of Social Determinants of Health Across a National Network of Community Health Centers. *Am J Prev Med* 2019;57:S65–73.
29 Goldstein BA, Bhavsar NA, Phelan M, *et al*. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *Am J Epidemiol* 2016;184:847–55.
30 Phelan M, Bhavsar NA, Goldstein BA. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *EGEMS (Wash DC)* 2017;5:22.
31 Matei A. On some reweighting schemes for nonignorable unit nonresponse. *Surv Stat* 2018;77:21–33. Available: http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_January_2018.pdf
32 Buttice MK, Highton B. How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? *Polit anal* 2013;21:449–67.