

Using Electronic Health Record Data for Public Health Surveillance of Diabetes among Young
Adults

Sarah Conderino

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Public Health
under the Executive Committee
of the Mailman School of Public Health

COLUMBIA UNIVERSITY

2023

© 2023

Sarah Conderino

All Rights Reserved

Abstract

Using Electronic Health Record Data for Public Health Surveillance of Diabetes among Young
Adults

Sarah Conderino

There is growing interest in using electronic health records (EHRs) for the surveillance of chronic diseases because these data contain a wealth of timely, clinical information on large samples of individuals. However, as these data are collected for clinical purposes, they may be prone to a number of biases that could affect their utility for public health practitioners or researchers. First, these data represent convenience samples of individuals who are in-care. These samples may not be representative of target populations for public health surveillance activities (e.g., the general population within a city or jurisdiction) by factors like demographics or health status, which could affect the generalizability of results. Second, key variables, including demographics or disease status, are often susceptible to missing data or misclassification, which could affect estimation of disease prevalence or risk factor associations. The goal of this integrated learning experience (ILE) was to assess the application of EHR data for chronic disease surveillance, focusing on the potential impact of selection and information biases for the case study of diabetes.

The first aim characterized the existing literature on defining diabetes status and type using EHRs from a population health perspective. The second aim externally validated diabetes prevalence estimates generated using EHR data from a large academic medical center in New York City (NYC) compared to traditional surveillance estimates from a local health survey. Various statistical methods, including raking, post-stratification, and multilevel regression with post-stratification, were applied to these real-world data and to simulated data to assess the ability to mitigate selection biases. Finally, the third aim externally validated EHR-based associations with potential diabetes risk factors (i.e., race/ethnicity and asthma) compared to estimates from national surveillance systems, including the Behavioral Risk Factor Surveillance System and National Health and Nutrition Examination Survey. Methods from the missing data and causal inference literature were then applied to assess the ability to control for misclassification of health outcomes in the EHR data.

Results from the literature review demonstrated that while there was no gold standard for defining diabetes using EHR data, definitions that prioritized sensitivity over specificity may be preferable for population health purposes. Based on this review, a flexible definition that searched for evidence of diabetes across diagnoses, medications, and lab results was used for the second and third aims. In the second aim, using statistical methods to account for demographic differences between the EHR sample and general population helped to remediate biases observed in the crude diabetes prevalence estimates. However, simulation results demonstrated that these methods may be insufficient when data are lacking for variables that are strong predictors of selection into the EHR sample. In the third aim, applying missing data or causal inference methods to control for misclassification of health outcomes greatly reduced the strength of the

association between asthma and diabetes status compared to naïve associations observed within the EHR sample, in alignment with observations from national health survey data.

Overall, the findings of this ILE suggest that naïve EHR analyses may yield biased estimates of diabetes prevalence or measures of association, driven in part by differences in healthcare utilization patterns across the population. However, applying epidemiologic frameworks can help control for and, importantly, characterize residual biases in these estimates. Future research is needed to assess the potential for selection and information biases across a variety of health outcomes, geographies, and EHR data sources to further inform the utility of these data for population health surveillance.

Table of Contents

List of Charts, Graphs, Illustrations.....	iii
Acknowledgments.....	vi
Dedication.....	vii
Chapter 1: Introduction.....	1
Chapter 2: Using Electronic Health Records for Population Health Surveillance of Diabetes: A Review of Computable Phenotype Definitions.....	4
2.1 Introduction.....	4
2.2 Methods.....	6
2.3 Results.....	6
2.3.1 Methods to Define Diabetes Computable Phenotypes.....	6
2.3.2 Internal Validity of Computable Phenotype Definitions	8
2.3.3 Transportability of Computable Phenotypes	10
2.4 Discussion.....	12
2.5 Conclusion	14
Chapter 3: Addressing Selection Biases within Electronic Health Record Data for Population Estimation of Diabetes Prevalence in New York City Young Adults.....	15
3.1 Introduction.....	15
3.2 Methods.....	18
3.2.1 Real-World Analysis Study Population.....	18
3.2.2 Prevalence Estimation & Performance	19

3.2.3 Simulation Analyses	20
3.3 Results.....	23
3.3.1 Real-World Analyses.....	23
3.3.2 Simulation Analyses	27
3.4 Discussion.....	30
3.5 Conclusion	35
 Chapter 4: Addressing Information Biases within Electronic Health Record Data for Examining Epidemiologic Associations with Diabetes Prevalence among Young Adults	 36
4.1 Introduction.....	36
4.2 Methods.....	40
4.2.1 Data Sources	40
4.2.2 Statistical Analyses	42
4.3 Results.....	44
4.4 Discussion.....	49
4.2 Conclusion	53
 Chapter 5: Conclusion.....	 54
References.....	57
Appendix A.....	65
Appendix B.....	70
Appendix C.....	77

List of Charts, Graphs, Illustrations

Main Text Tables

Table 1: American Diabetes Association’s Criteria for the Diagnosis of Diabetes.....	8
Table 2: Demographic Profile of the NYU Langone EHR Sample and NYC General Population, Young Adults Aged 18-44 Years.....	24
Table 3: Diabetes Prevalence among NYC Young Adults 18-44 Years, Estimated from the NYU Langone Health Electronic Health Record vs. NYC Community Health Survey (NYC CHS) Gold Standard.	24
Table 4. Descriptive Summary of NYU Patient Population by Complete Case Status.....	45
Table 5: Descriptive Summary of NYU Patient Population by Diabetes Status.	46

Main Text Figures

Figure 1: Data Simulation Directed Acyclic Graph with Baseline Odds Ratio (OR) Associations.	21
Figure 2: Characterization 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 Neighborhood.	26
Figure 3: Mean Relative Bias in the EHR-Based Estimates vs. True Diabetes Prevalence by Simulation Scenario.	29

Figure 4: Directed Acyclic Graphs (DAGS) for Measurement Error in Exposure and Outcome Variables	39
Figure 5: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, EHR-Based Estimates vs. Health Survey Estimates	48

Appendix Tables

Appendix Table 1: Literature Review Results for Diabetes Computable Phenotype Definitions.	65
Appendix Table 2: Demographic Profile of the NYU Sample and General Population under Different Catchment Area Definitions, NYC Young Adults Aged 18-44 Years.	72
Appendix Table 3: Overall Diabetes Prevalence Estimates (and 95% CIs) under Different Catchment Area Definitions, NYC Young Adults Aged 18-44 Years.	73
Appendix Table 4: Relative Difference in EHR-Based Diabetes Prevalence Estimates from Gold Standard under Different Catchment Area Definitions, NYC Young Adults Aged 18-44 Years.	74
Appendix Table 5: Coverage in Overall EHR-Based Estimates by Adjustment Method and Simulation Scenario.	76
Appendix Table 6: Case-Insensitive Search Terms for Endocrinology Review of Systems.....	77
Appendix Table 7: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, Health Survey Estimates.	78
Appendix Table 8: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, EHR-Based Estimates	78
Appendix Table 9: Sensitivity Analyses of Complete Case Definitions: Descriptive Summary of NYU Patient Population by Diabetes Status.....	79

Appendix Table 10: Sensitivity Analyses of Complete Case Definitions: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, EHR-Based Estimates Complete Case and Missing Data Estimates 80

Appendix Figures

Appendix Figure 1: Relative Bias in the Neighborhood-Level EHR-Based Estimates vs. the True Diabetes Prevalence by Simulation Scenario. 75

Appendix Figure 2: Hypothesized Directed Acyclic Graph for Information Bias and Associations between Asthma and Diabetes. 77

Acknowledgments

I would like to extend my deepest gratitude to my committee members, Dr. Sandra Albrecht, Dr. Lorna E. Thorpe, Dr. Rebecca Anthopolos, Dr. Shannon Farley, and Dr. Catherine Stayton, for their insightful comments and mentorship during this Integrated Learning Experience. I am also grateful to my academic advisor, Dr. Mary Ann Chiasson, and academic director, Liliane Zaretsky, for their support throughout the Doctor of Public Health Program. I would also like to thank the New York City Department of Health and Mental Hygiene, who provided New York City Community Health Survey data for this work.

Dedication

This dissertation is dedicated to Johanna Finn and Margaret Sorensen.

Chapter 1: Introduction

Population health surveillance is an essential function of public health practice. Comparing disease burden across places, demographic groups, and time periods helps governments and public health organizations understand trends and clusters, evaluate the effectiveness of interventions, measure health equity, and distribute resources to the populations of highest risk. Yet many jurisdictions rely on outdated surveillance data sources, such as health surveys or administrative claims, which can be resource-intensive and have limited internal or external validity.¹⁻⁶ The widespread adoption of electronic health records (EHRs) poses a strategic opportunity to advance surveillance and epidemiologic practices due to their large sample sizes, near real-time availability, and wealth of longitudinal clinical data.⁷⁻⁹

While EHR data have strengths that could benefit public health practices, there are a number of challenges inherent to these data. First, EHRs represent convenience samples that are restricted to individuals who are in-care. These data are therefore susceptible to selection biases.^{10,11} Patient populations, particularly those from private or academic medical centers, may be non-representative of the general population by demographic factors such as age, race/ethnicity, socioeconomic status (SES), immigration status, and sex,¹² which are key determinants of health. Patients within healthcare systems are also typically sicker than the general population, which could contribute to overestimation of disease burden when using conventional statistical methods.¹³

Second, classification of key variables, including demographics or disease status, may be susceptible to information biases.^{11,14} Since EHRs are designed for medical purposes, variables that are not perceived to be clinically relevant, such as race/ethnicity, may have high levels of

missing data or poor quality.¹⁵ For the classification of disease status, non-differential misclassification could occur from simple data entry errors, while differential misclassification could occur if certain diagnosis codes are preferentially used based on differences in reimbursement rates.¹¹ A diagnostic suspicion or surveillance bias could occur if certain at-risk subgroups have increased screening for health factors.¹¹ For example, patients who are obese may be more likely to be screened for diabetes with hemoglobin A1c testing, or patients who visit the health system more frequently may have more opportunities to receive a diagnosis. In addition, when analyzing EHR data from a single healthcare system, pertinent information may be missing due to patients receiving care across multiple distinct healthcare systems.

The overarching goal of this integrated learning experience (ILE) is to explore biases in EHR data and methods for addressing these biases in order to increase the utility of EHR data for public health surveillance of chronic conditions. This ILE will focus on the case study of diabetes, a serious, chronic condition that is increasingly prevalent in the United States,^{16,17} and thus would benefit from innovations to provide timely, longitudinal surveillance and epidemiologic data to allow public health and healthcare agencies to respond to the rapidly changing landscape of burden and treatment. Specifically, the aims of this ILE are (1) to summarize the current literature, including gaps and opportunities, on using EHRs for the classification of diabetes status and type, (2) to explore the impact of selection bias on diabetes prevalence estimation, and (3) to explore the impact of information bias on the estimation of risk factor associations with diabetes.

Chapter 2 reviews the existing literature on defining diabetes status and type using EHRs from a population health perspective. Chapter 3 explores the impact of selection biases on the estimation of diabetes prevalence using real-world EHR data from NYU Langone Health.

Various statistical methods are applied to both these real-world data and to simulated data to assess the ability to correct for selection biases. Chapter 4 explores the impact of information biases on the estimation of risk factor associations with diabetes using the NYU Langone EHR data. Methods from the missing data and causal inference literature are applied to assess the ability to correct for information biases. Finally, chapter 5 provides the overall conclusions of the ILE and recommendations for future public health research on chronic disease surveillance using EHRs.

Chapter 2: Using Electronic Health Records for Population Health

Surveillance of Diabetes: A Review of Computable Phenotype

Definitions

2.1 Introduction

Diabetes is a serious, chronic condition that affects an estimated 37 million children and adults within the United States (US) and is increasing in both prevalence and incidence.¹⁸⁻²⁰ Evidence also suggests that disparities in diabetes burden by socioeconomic status, race/ethnicity, and geography are widening over time.^{19,21,22} Timely and accurate surveillance data on trends and patterns in diabetes prevalence, including both Type 1 and Type 2, are essential for routine public health practice activities, such as evaluating the effectiveness of interventions, measuring health equity, and distributing resources to populations of highest risk. However, surveillance data for chronic diseases like diabetes are typically informed by telephone- or address-based health surveys (e.g., the Behavioral Risk Factor Surveillance System (BRFSS), National Health and Nutrition Examination Survey (NHANES), or National Health Interview Survey (NHIS)), which have a number of limitations – financial expense is high,³ validity of self-reported disease classification is low,^{5,23,24} and delays between data collection and dissemination can be long.¹ Exacerbating this, response rates to health surveys have decreased over time, threatening the generalizability of the survey results,² and many surveys are limited to national or state estimation, limiting their transportability to smaller geographic scales or underrepresented populations.

The widespread adoption of electronic health records (EHRs) among US hospitals and office-based practices since 2014²⁵ poses a strategic opportunity to advance surveillance and

epidemiologic practices. Benefits of these data are numerous. Data can be available in near real-time, allowing for more timely monitoring of trends and patterns in disease burden. Large sample sizes can also provide precise estimates for small geographic areas. In addition, EHRs provide a rich source of longitudinal, clinical data, such as physical measurements and lab results, which are otherwise costly and challenging to obtain.⁷⁻⁹ Recently, there has been increasing interest among researchers and public health practitioners on how EHR data can be leveraged for population health purposes.²⁶⁻²⁸ Moreover, large, national research networks including the Centers for Disease Control and Prevention-funded Multi-state EHR-based Network for Disease Surveillance (MENDS) and Diabetes in Children and Young Adults (DiCAYA) networks, have been established to explore how EHRs can be used for chronic disease, and specifically diabetes, surveillance.^{29,30}

Historically, efforts to define or classify patients with diabetes from EHR data come from comparative effectiveness or epidemiologic research studies.³¹⁻³⁴ While these studies have generally defined diabetes status within patient-based study populations, their definitions and findings could help inform appropriate definitions for population health surveillance purposes. A smaller subset of EHR-based studies have defined diabetes explicitly for surveillance purposes.³⁵⁻³⁷ However, definitions may vary based on target populations or the type of EHR data used, such as data from a single healthcare institution versus data from multiple health care institutions that have been linked through a health information exchange (HIE) or clinical research network (CRN). The goal of this review is to evaluate how EHR data can be used to define diabetes status and type among both children and adults from a population health surveillance perspective. In this paper, we summarize methods used, internal validity, and transportability of various EHR-based diabetes definitions from the literature. We then assess

how these definitions would translate for population health purposes and provide recommendations for future EHR-based diabetes surveillance research.

2.2 Methods

This review included studies from 2012-2021 from the US or Canada that defined diabetes status among children or adults using “computable phenotypes”, which are models or algorithms that define a condition using data that is solely processed by a computer (n=18).³⁸ There is no established gold standard for classifying diabetes status or type using EHRs.³⁹

2.3 Results

2.3.1 Methods to Define Diabetes Computable Phenotypes

Almost all studies identified in this review included “rule-based” definitions, or computable phenotypes based on pre-specified logic-based inclusion and exclusion criteria (Appendix A Table 1) (n=17).⁴⁰ Rule-based computable phenotypes are often informed by clinical guidelines or expertise, such as the American Diabetes Association’s guidelines for diagnosing and classifying diabetes (Table 1).^{31,41-43} For example, many studies (n=14) classified individuals’ diabetes status based on some combination of the following criteria that could be pulled from lab results, diagnosis codes, and medications:^{39,44,45}

- Elevated hemoglobin A1C \geq 6.5%
- Elevated random blood glucose \geq 200 mg/dL
- Elevated fasting blood glucose \geq 126 mg/dL
- Inpatient or outpatient diabetes diagnosis codes
- Anti-diabetic medications

While most studies allowed these criteria to occur at any point within a given time window (e.g., past two years), few studies had specific temporal guidelines between certain criteria (e.g., anti-diabetic medication within 90 days of an elevated A1c lab result) (n=2).

Studies commonly defined rule-based criteria using structured data elements, such as International Classification of Disease (ICD) diagnosis codes, Logical Observation Identifiers Names and Codes (LOINC), RxNorm codes, and National Drug Codes (NDCs).^{31,44} While most studies explicitly specified the lists of ICD diagnosis codes underlying their definitions,^{37,39,46} methods documenting how researchers identified the diabetes-related labs or medications were frequently lacking. Publishing these lists of codes or methods in online repositories, such as the National Library of Medicine's Value Set Authority Center (VSAC) or the Phenotype KnowledgeBase (PheKB), could facilitate replication of computable phenotypes across studies.^{47,48}

Some of the identified studies (n=3) also incorporated natural language processing (NLP) to pull information from text-based variables, such as through keyword searches of semi-structured medication or lab test names (Appendix A Table 1).^{48,49} NLP methods were also used on free-text fields, such as consult notes, discharge summaries, or admission notes. These studies identified diabetes using keyword searches while incorporating more advanced logic to exclude negations (e.g. "no diabetes") or family history of diabetes (e.g., "mother has diabetes").^{32,50} NLP methods could improve upon the flexibility and transportability of computable phenotypes as compared to using lists of diagnosis, lab, or medication codes, which can be institution-specific or change over time. However, since these methods are also susceptible to the uniformity, accuracy, and completeness of text-based fields, NLP may best serve to augment, rather than replace, more traditional queries of structured data elements, as used in the studies in this review.^{32,49,50}

Data-driven approaches, including machine learning or regression-based methods, to automatically classify diabetes status or type without pre-specification of the inclusion or

exclusion criteria were less commonly used (Appendix A Table 1) (n=3). When these methods have been used, they have yielded mixed results as compared to more traditional rule-based approaches. Researchers who have tested classification and regression tree (CART) or multinomial logistic regression to predict diabetes status and diabetes type among children observed that these methods did not outperform traditional rule-based approaches.^{37,51} Limitations of these approaches include decreased interpretability of findings, increased complexity in implementation, and increased potential of overfitting algorithms to the data, leading to poor generalizability or transportability to other external populations of interest.

Table 1: American Diabetes Association’s Criteria for the Diagnosis of Diabetes.

Fasting plasma glucose (FPG) \geq 126 mg/dL (7.0 mmol/L). Fasting is defined as no caloric intake for at least 8 hours ^a
2-hour plasma glucose (PG) \geq 200 mg/dL (11.1 mmol/L) during oral glucose tolerance test (OGTT)
A1C \geq 6.5% (48 mmol/mol)
In a patient with classic symptoms of hyperglycemia or hyperglycemic crisis, a random plasma glucose \geq 200 mg/dL (11.1 mmol/L)

^a *In the absence of unequivocal hyperglycemia, diagnosis requires two abnormal test results from the same sample or in two separate test samples.*⁴³

Note: Reprinted from Committee ADAPP. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022. Diabetes care. 2021;45(Supplement_1):S17-S38.

2.3.2 Internal Validity of Computable Phenotype Definitions

The internal validity or performance of computable phenotype definitions was often contrasted to gold standard findings from manual medical chart review (Appendix A Table 1) (n=15). Common performance measures included sensitivity (the proportion of true diabetes cases who were classified as diabetic through the computable phenotype), specificity (the proportion of true non-diabetes cases who were classified as non-diabetic), or positive predictive value (PPV – the proportion of those classified as diabetic who truly had diabetes). These performance measures are dependent on one another; as specificity increases, sensitivity decreases and PPV increases.⁵² Since researchers have to balance these performance measures,

the ultimate choice or recommendation for a computable phenotype definition may depend upon the purpose of the study.

Traditional epidemiologic or comparative effectiveness research studies rely on correctly identifying samples of cases and controls or cohorts of disease-free individuals at baseline. Therefore, these types of studies may prioritize computable phenotype definitions with higher specificity or PPV when creating their study samples. Definitions that require evidence of diabetes within multiple sources of clinical elements (e.g., ICD codes *and* laboratory results)^{31,39,50} are more likely to limit the number of individuals who are falsely classified as diabetic, thus improving specificity or PPV. Researchers using EHR data from Vanderbilt University Medical Center, for example, found that definitions for type 1 and type 2 diabetes using ICD codes alone produced PPVs of 59% and 65% respectively. Definitions that required evidence within at least two clinical sources (including ICD codes, primary notes, or medications) had PPVs of 91% and 81% for type 1 and type 2 diabetes respectively.⁵⁰ While highly specific definitions may be required for rigorous epidemiologic or comparative effectiveness research, they may lack the sensitivity desired for population health surveillance.

In contrast, broad definitions that search for any evidence of diabetes across multiple sources of clinical elements (e.g., ICD codes *or* laboratory results) may improve upon the sensitivity, or ability to detect patients with diabetes.^{44,53,54} In a study leveraging a network of ambulatory practices in New York City, researchers could identify 87% of true diabetes cases using ICD codes alone, but could identify 94% of true cases when they modified their definition to also include elevated hemoglobin A1C (HbA1c) results or diabetes-related medications.⁵³ Studies that are focused on population health surveillance often hope to leverage EHR data to estimate disease prevalence within in-care populations. These types of studies may prioritize

broader computable phenotype definitions with higher sensitivity, particularly since EHRs have the known limitation of missing or incomplete data that will hinder the ability to correctly identify all diseased individuals. However, within diabetes research, there is often the added complexity of distinguishing diabetes types, and definitions that are too broad could result in the misclassification of those with type 1 diabetes as type 2.⁵⁴

2.3.3 Transportability of Computable Phenotypes

A number of factors can affect transportability of computable phenotypes definitions across EHR data sources, time, or populations. As discussed previously, definitions may be built using standardized codes that could change over time or be institution-specific, and providers may have differential documentation practices that are influenced by billing or institutional policies.^{31,55,56} For example, a computable phenotype definition based on outpatient anti-diabetes medications correctly identified 89% of true diabetic patients at the University of North Carolina Health Care System (UNC) but only 13% of true diabetic patients at the Medical University of South Carolina (MUSC).³⁷ Only 8% of sampled MUSC patients were identified as meeting this computable phenotype as compared to 49% of sampled UNC patients, perhaps suggesting the use of distinct institutional methods for coding or storing medication data that were not consistently captured by the algorithm. Similarly, researchers at Kaiser Permanente Colorado (KPCO) recommended removing the urine acetone test strip criterion from the Klompas type 1 diabetes computable phenotype.⁴⁵ This specific criterion originally performed with a PPV of 100% within the Atrius Health ambulatory practice network⁴⁴ but performed much worse within the KPCO data, with researchers noting that the dispensing of these strips was low within their system.⁴⁵

Moreover, the characteristics of the EHR data source can affect the transportability of computable phenotype definitions. Diabetes computable phenotypes may have better performance within healthcare systems that are more likely to serve as the primary source of diabetes care, such as those that include large ambulatory networks or integrated care organizations (e.g., Kaiser Permanente or the Veterans Health Administration). In addition, EHR data from single healthcare institutions likely provide an incomplete record of patients' care received, particularly in dense urban areas where individuals have access to and may receive care from a range of different institutions. In contrast, EHR data from integrated care organizations or from EHR networks that link patient care across contributing institutions (e.g., PCORnet CRNs) will provide a more accurate representation of patients' medical histories. A commonly used computable phenotype for distinguishing diabetes type in children was based on whether the patient had a majority of type 1 versus type 2 diagnosis codes. When this definition was applied at independent academic healthcare institutions, it achieved PPVs for type 2 diabetes ranging from 55% to 72%.^{51,57} However, when this definition was applied to an integrated managed care organization, it achieved a PPV of 89%.⁴⁶ The more complete information afforded by this managed care organization likely contributed the accuracy of the type 2 classification for this computable phenotype.

Transportability of computable phenotypes is also dependent on the population under study, and PPV generally increases as prevalence of the outcome increases.^{46,52} For diabetes research, the age of the population of interest is particularly important, as type 1 diabetes is more prevalent within children while type 2 diabetes is more prevalent within adults. A given computable phenotype for distinguishing type 1 diabetes may therefore perform better within younger populations, while a given computable phenotype for distinguishing type 2 diabetes may

perform better within older populations. This was illustrated when researchers applied the Klompas algorithm, originally designed within a target population of all ages,⁴⁴ to a patient population of children aged 10 to 18 years.⁵⁸ In the original study, the computable phenotype to distinguish type 1 and type 2 diabetes performed with a sensitivity and PPV of 65% and 88% respectively for type 1 diabetes and 100% and 95% respectively for type 2 diabetes.⁴⁴ When this computable phenotype was applied to children, its performance improved for the classification of type 1 diabetes (sensitivity of 99% and PPV of 91%) and worsened for the classification of type 2 diabetes (sensitivity of 77% and PPV of 87%).⁵⁸

2.4 Discussion

Based on this review, computable phenotypes for population health surveillance purposes should be appropriately broad, incorporating information from diagnosis codes, lab results, and medications or dispensings. This will help improve sensitivity, or the ability to capture true diabetes cases, within imperfect EHR data. Additionally, *optional* criteria that include evidence of diabetes across multiple sources or time points (e.g., elevated A1c on two separate dates OR a diabetes diagnosis code) may help to prevent the false classification of individuals with prediabetes or single aberrant lab results. However, definitions should avoid *requiring* documentation of evidence across multiple sources or time points (e.g., a clinically consistent definition that requires two positive lab tests on separate dates, Table 1), as this may be too restrictive for patients who do not receive regular healthcare. Characterizing patterns in diabetes prevalence is a key function of population health surveillance, and these restrictive definitions could result in systematic biases that distort our understanding of social determinants of health. For example, lower-income individuals may have fewer interactions within a healthcare system,

and definitions that require a level of consistent care could systematically underestimate diabetes prevalence within these populations.

Computable phenotypes for population health surveillance purposes should also be flexible enough to allow for transportability across populations, time, and organizations. Researchers should allow for differences in definitions by age group, both because the relative burden of type 1 diabetes differs among children and adults and because healthcare utilization differs across the lifespan. Children or older adults have more regular and greater continuity of care,^{59,60} which could increase their likelihood of receiving a diabetes diagnosis as compared to younger adults.¹¹ As trends and patterns in diabetes are changing, research is needed to continue to monitor the performance of computable phenotypes over time. Since many definitions are also institution-dependent, more effort is needed to streamline replicability across organizations, such as through publishing codes in online repositories and moving towards EHR common data models. Additional research is also needed to explore the impacts of and ability to address misclassification due to factors like patient demographics (e.g., using uniform A1c criteria that are insensitive to racial differences for glucose-defined diabetes)^{61,62}, provider non-adherence to clinical screening guidelines, missing data, and duration of longitudinal records.^{31,37,63,64}

With any population health surveillance work using EHRs, researchers must acknowledge key limitations with these data. Importantly, these data are collected on individuals who are in-care and may only be generalizable to these populations. Almost one-quarter of adults with diabetes are undiagnosed, with higher prevalence among communities of color.⁶⁵ EHR data may therefore fail to capture a considerable number of diabetes cases within populations who are undiagnosed or do not receive adequate care. In addition, the proprietary and identifiable nature of these data often translate to increased costs or inaccessibility for public health researchers.⁶⁶

Despite these limitations, EHR data offer a number of strengths, including timeliness, geographic granularity, and clinically-based measurements, and when accessible, can complement data from existing surveillance systems.

2.5 Conclusion

A robust literature base has emerged in the past decade to inform computable phenotype definitions for defining diabetes status and type among children and adults, which includes diverse purposes, ranging from epidemiologic and comparative effectiveness studies to population health surveillance. In general, rule-based approaches for these definitions have many advantages, including increased transportability and interpretability through their grounding in clinical standards. However, augmenting these definitions to also include more advanced methods, such as natural language processing, could improve computable phenotype flexibility to handle changes across time or institutions. When selecting or developing a novel diabetes computable phenotype, researchers should consider their study purpose, data sources, and target population of inference. For population health surveillance purposes, broader definitions that search for evidence of diabetes across multiple clinical sources may be preferable, particularly when using EHR data from a single healthcare institution. In addition, surveillance will likely require different computable phenotypes to detect and distinguish diabetes type within children than within adults.

Chapter 3: Addressing Selection Biases within Electronic Health Record Data for Population Estimation of Diabetes Prevalence in New York City Young Adults

3.1 Introduction

Increasingly, public health researchers and practitioners have explored how electronic health records (EHRs), which offer near real-time clinical data on large patient populations, can be leveraged for valid and reliable public health surveillance purposes.^{67,68} While EHRs offer a compelling opportunity for surveillance, patient populations, particularly those from private or academic medical centers, may be non-representative of the general population with respect to demographic characteristics, such as age, sex, race/ethnicity, and socioeconomic status (SES).¹² From a health status perspective, patients represented within EHR data are typically sicker than the general population.¹³ This can contribute to overestimation of disease prevalence and incidence for public health surveillance purposes, where the target population for inference is typically the general population, including those not in-care. Thus, differences between the EHR sample and the general population introduces the potential for selection bias in EHR-based surveillance.¹¹

Data collected in EHRs represent a nonprobability sample in which individuals do not select into the health system at random. Moreover, the process by which individuals select into the sample is unknown. The missing data lexicon has been applied to nonprobability samples to characterize different mechanisms by which individuals may be included within these data.⁶⁹ Akin to missing at random, selection at random (SAR) describes scenarios whereby the probability of selection depends on some observed characteristics of the individuals, but given

those characteristics, is independent of unobserved outcomes from individuals excluded from the sample.^{70,71} In contrast, missing not a random, herein called selection not at random (SNAR), represents nonignorable selection processes whereby the probability of selection is dependent on unobserved outcomes, even after adjusting for observed covariates.^{70,71}

In practice, statistical methods from the nonprobability literature, including raking, post-stratification, and multilevel regression with post-stratification (MLRP), have previously been applied to EHR samples to estimate prevalence of a variety of diseases, including diabetes, within wider geographies or populations.^{35,72,73} These methods often rely on an assumption of SAR, controlling for covariates that are measured in both the EHR sample and population, such as basic demographics or neighborhood proxies of SES. These methods are appealing as they are relatively easy to implement, have demonstrated effectiveness in the political science literature,⁷⁴⁻⁷⁶ and given that the few statistical methods that allow SNAR rely on untestable model assumptions without further data collection.⁷⁴⁻⁷⁷ However, the tendency for EHRs to over-represent sicker individuals increases the plausibility of SNAR, suggesting that residual selection biases may remain within prevalence estimates generated using these more common methods.

One approach used to assess for residual biases in EHR-derived surveillance estimates is to compare to “gold standard” estimates, such as those derived from population representative health surveys with known sampling weights.^{35,72,73} Although the comparison with gold standards helps to externally validate EHR-derived estimates, this does not facilitate understanding the conditions under which different methods will provide valid estimation. For example, Chen et al., used data from MDPHnet, a network including EHR data from three large health systems in Massachusetts (MA), to compare MLRP-adjusted prevalence estimates for five health outcomes to estimates from the Behavioral Risk Factor Surveillance System 500 Cities

project.⁷² Investigators found that the external validity of the EHR-based estimates was variable by both the health outcome of interest and by the individual health system, consistent with other research.^{35,72} These differences could arise from a variety of factors, such as unmeasured confounders or selection processes that operate in specific populations or diseases. Moreover, not all jurisdictions will have gold standard estimates available, and even if gold standard estimates are available for a geographic area, they may not be readily available for subgroups defined by demographics or neighborhoods. Data simulations provide the opportunity to test these methods under controlled conditions, which may inform the transportability of these methods to other such populations.

As part of wider efforts to estimate diabetes prevalence among young adults using EHR data,²⁹ we conducted a multi-step process to evaluate common bias adjustment methods (raking, post-stratification, and MLRP), first in a “real-world” setting where we could evaluate validity against “gold-standard” estimates, and second, in simulation settings where we could manipulate selection processes and evaluate validity against the controlled truth. Real-world analyses were conducted using EHR data from a medical center in New York City (NYC), a jurisdiction with granular, high-quality gold standard data sources from external surveillance systems. NYC is also home to several academic medical centers and a large system of 11 public hospitals; thus, the likelihood of one private health system being representative of the general population is low. Based on the results of this initial real-world analysis, we then ran a series of simulations applying bias adjustment methods to data generated under various assumptions on the underlying selection mechanisms (e.g., selection dependent on unmeasured factors or on diabetes status). Simulations reflected our hypothesized factors that could have contributed to residual biases observed within the initial real-world analyses. The overarching goal of the paper was to

compare these bias adjustment methods under real-world and simulation settings to help inform the broader discussion on how to effectively use EHRs for population-level surveillance purposes.

3.2 Methods

3.2.1 Real-World Analysis Study Population

NYU Langone Health (NYU) is a large academic medical center that serves patients throughout NYC but has primary service areas in the boroughs of Manhattan, Brooklyn, and Queens. Longitudinal NYU EHR data were obtained for all NYC-resident patients aged 18-44 years who had an inpatient or outpatient encounter from 2017-2019. As some researchers limit EHR samples to health system “catchment areas” to better reflect primary populations served by their facilities and to potentially reduce selection biases,²⁹ we conducted sensitivity analyses varying the resident inclusion criteria to restrict to NYC neighborhoods within different definitions of NYU “catchment areas” (Appendix B). Main analyses included all NYC residents since prevalence estimation to the full jurisdiction is of greater public health relevance.

Patients were classified using five demographic variables: age group (18-29 and 30-44 years), sex (male or female), race/ethnicity (White, Black, Latino, and Asian/Other), Medicaid insurance status (yes vs. no), and neighborhood of residence (Public Use Microdata Areas, PUMAs, n = 55). Race/ethnicity was imputed for those with unknown race/ethnicity (19%) using the Bayesian Improved Surname Geocoding (BISG) methods.⁷⁸ All patients with an unknown/other age or sex were excluded (<1%). Patients with diabetes were classified following methods proposed by Avramovic, et al., as those with at least two diagnoses for diabetes during clinical encounters, one diagnosis and at least two elevated A1C lab results $\geq 6.5\%$, or at least one anti-diabetes prescription medication (not including metformin or acarbose).⁷⁹

3.2.2 Prevalence Estimation & Performance

We calculated diabetes prevalence for NYC overall and by neighborhood within the crude EHR sample (i.e., the percentage of patients who were classified as having diabetes) and adjusted to the general population of NYC adults aged 18-44 years. We defined equivalent general population demographic variables using American Community Survey (ACS) 2019 5-year data obtained through IPUMS USA, a line-level sample of ACS data.⁸⁰ Adjusted prevalence estimates were calculated using three approaches: raking, post-stratification, and MLRP.

Raking and post-stratification prevalence estimates and confidence intervals were calculated using the “survey” R package,⁸¹ adjusting for the five demographic variables. MLRP estimates were calculated following previously published methods.⁸² First, a multilevel logistic regression model was fit including fixed effects for binary demographics and random effects for all non-binary individual-level demographics. Full details on the final model specification and sensitivity analyses of alternative model specifications that include neighborhood-level social determinant of health (SDOH) and health outcomes are found in Appendix B. Model predicted probabilities were applied to the post-stratification weights for the equivalent covariates within the general population. Confidence intervals were calculated using parametric bootstrapping with the “lme4” R package.⁸³

The gold standard prevalence estimates for comparison were calculated using pooled 2015-2020 data from the population-representative NYC Community Health Survey (CHS). The NYC CHS is a cross-sectional telephone survey conducted by the NYC Department of Health and Mental Hygiene. The survey includes a variety of health topics, including diabetes and other chronic conditions, and is conducted annually on a stratified random sample of approximately 10,000 NYC adults.⁸⁴ NYC CHS prevalence estimates and confidence intervals were calculated

using the “survey” R package following published methods.⁸⁴ 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; (2) statistical equivalence to the gold standard estimate through the two one-sided test (TOST) test using an alpha of 0.05 and equivalence bounds of 0.005; and (3) the Pearson correlation coefficient between the neighborhood-level EHR and gold standard estimates.

3.2.3 Simulation Analyses

Based on the results from the real-world analysis, we hypothesized that residual biases could be present within the EHR-derived prevalence estimates due to our inclusion of patients outside of the NYU catchment area or due to unaddressed selection dependencies, such a selection not at random mechanism (i.e., dependent on diabetes status). Simulations were run to probe the extent to which residual biases remain under three scenarios: (1) sample inclusion criteria depends on the definition of the health system catchment area (e.g., excluding neighborhoods that have a lower proportion of total residents captured within the sample); (2) selection is dependent on an unmeasured factor, for which there is proxy/auxiliary information measured in the sample and general population (e.g., those with high SES have higher odds of being within the EHR sample but individual SES is measured with an imperfect proxy of insurance type); and (3) selection is dependent on diabetes status, which is less likely to be measured for those not within the sample (e.g., those with diabetes have higher odds of being within the EHR sample).

Simulated populations were composed of 500,000 individuals equally distributed across 50 neighborhoods. The hypothetical health facility was defined to be located in a central neighborhood with the remaining neighborhoods classified in a 7x7 matrix with one to three-unit

distance measured from the facility. Individuals' demographics were classified for age (18-29 or 30-44 years), race/ethnicity (White, Black, Latino, and Asian/Other), and sex (male or female).

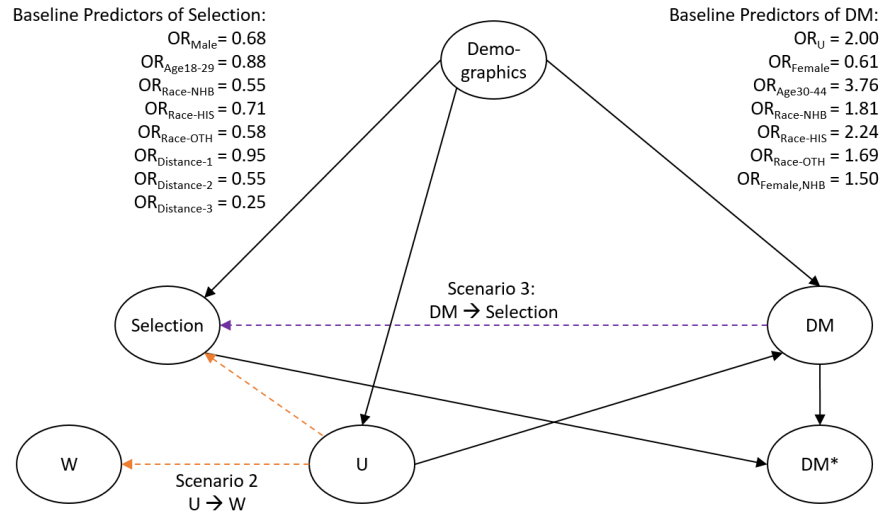


Figure 1: Data Simulation Directed Acyclic Graph with Baseline Odds Ratio (OR) Associations.¹

Simulations defined individuals' diabetes status and selection into the EHR sample based on assumed associations, as displayed in Figure 1, that were informed through real-world analyses. Diabetes ("DM") was defined using a probability function based on fixed effects with demographic variables (as informed through CHS 2019-2020 data) and on random effects to generate heterogeneity in diabetes prevalence across neighborhoods. Selection into the hypothetical EHR sample was also defined using a probability function based on fixed effects with demographics (as informed through an analysis of NYU Langone EHR data), neighborhood distance from the facility, and on random effects to generate heterogeneity in selection for the

¹ Observed diabetes within those selected into the EHR sample; Scenario 2 (orange): modified the level of misclassification of the auxiliary variable W compared to the unobserved variable U at levels equivalent to 10%, 30%, 50%, 70%, and 90% misclassification; Scenario 3 (purple): modified the association between diabetes and selection at OR levels of 0.33, 0.67, 1.0, 1.5, and 3.0.

interaction of sex and race/ethnicity. Baseline associations between all variables, as measured using odds ratios (ORs), 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 EHR sample of 10%.

Simulation scenario 1 tested five sample inclusion criteria to define catchment areas based on incremental quantiles of the proportion of the total neighborhood population captured in the sample (ranging from including all neighborhoods to only neighborhoods within the top quantile of proportion of the population captured). Simulation scenario 2 introduced a binary individual-level, unobserved variable “U” that was associated with DM (OR=2.0) and selection (OR=0.7), which was modeled after observed patterns with household poverty level.⁸⁴ 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 2, U was not included in the adjustment procedures but W was included. Finally, simulation scenario 3 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. These OR levels were chosen to assess moderate- and high-strength SNAR mechanisms operating in both directions (those with DM have lower ($OR < 1$) or higher ($OR > 1$) odds of selection) compared to an SAR mechanism ($OR = 1.0$). For each scenario, 100 simulations were run.

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 were generated at the overall and neighborhood level for each simulation. In scenario 1, raking and post-stratification did not include neighborhood in the adjustment procedures since the sample did not include all neighborhoods,⁸⁵ and MLRP was

performed both with and without a random effect for neighborhood. In scenarios 2 and 3, the sample was not restricted based on neighborhood, therefore all adjustment procedures included covariates for demographics and neighborhood.

We assessed performance of each adjustment method using two measures: (1) relative bias, defined as the average percent difference between the true diabetes prevalence in the full population and the estimated diabetes prevalence within the sample; and (2) coverage probability of the 95% CI, defined as the percentage of simulations with a true diabetes prevalence falling within the 95% CI. All analyses were performed using R version 4.1.2.

3.3 Results

3.3.1 Real-World Analyses

A total of 454,612 young adults were identified in the NYU EHR sample. Compared to the NYC general population, the EHR sample had overrepresentation of White (1.6-fold), female (1.2-fold), and older-aged (1.1-fold) individuals (Table 2). Representation varied more substantially across the 55 NYC neighborhoods, reflecting increased capture of the general population within neighborhoods surrounding the NYU hospitals in lower Manhattan and Brooklyn (Figure 2A).

The gold standard diabetes prevalence among young adults was 3.33% (95% CI: 3.02-3.67). A total of 3.09% of the young adult EHR sample were classified as having diabetes (95% CI: 3.04-3.14), representing a significant 0.24 percentage point decrease or a -7.88% relative difference from the gold standard estimate (Table 3). Adjusted prevalence estimates using raking (3.55%), post-stratification (3.54%), and MLRP (3.55%) were all significantly higher than the crude EHR-based prevalence estimate. Adjusted estimates were also higher than the gold

standard, with positive relative differences of approximately 6%, but were statistically equivalent at the equivalence bound of 0.005.

Table 2: Demographic Profile of the NYU Langone EHR Sample and NYC General Population, Young Adults Aged 18-44 Years.

	NYC General Population ^a	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%

^aDefined using American Community Survey (ACS) 2019 5-year data obtained through IPUMS USA.⁸⁰

Table 3: Diabetes Prevalence among NYC Young Adults 18-44 Years, Estimated from the NYU Langone Health Electronic Health Record vs. NYC Community Health Survey (NYC CHS) Gold Standard.

	Prevalence (%) (95% CI)	Relative Difference from Gold Standard (NYC CHS) ^a
Gold Standard		
NYC CHS	3.33% (3.02-3.67)	-
EHR-Based		
Crude	3.09% (3.04-3.14)	-7.88%
Raking	3.55% (3.46-3.63)	6.02%*
Post-stratification	3.54% (3.43-3.64)	5.75%*
MLRP	3.55% (3.47-3.63)	6.16%*

*Statistically Equivalent to the gold standard through the TOST test $\alpha = 0.05$, equivalence bounds = 0.005.

^aPercent difference from the gold standard estimate, the New York City Community Health Survey.

At the neighborhood-level, there was moderate, significant correlation ($R = 0.5$) between EHR-based and gold standard prevalence estimates; though, as with the overall estimates, neighborhood-level EHR estimates were generally higher than the neighborhood-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 2D).

Sensitivity analyses varying the residential inclusion criteria to NYU catchment areas found that the crude EHR-based prevalence estimate decreased when restricting to neighborhoods where a greater proportion of the general population was captured in the sample (Appendix B Table 2). When these restricted samples were adjusted and externally validated to the *full NYC general population*, they produced estimates that were significantly equivalent and closer to the gold standard. However, when these estimates were adjusted and externally validated to *the catchment area general population* (e.g., NYC residents within the catchment area neighborhoods), where differences in demographic representativeness were less pronounced, they produced non-statistically equivalent estimates that were higher than the gold standard. Sensitivity analyses including neighborhood-level SDOH and health outcomes in the MLRP model did not meaningfully affect the overall or neighborhood-level prevalence estimates.

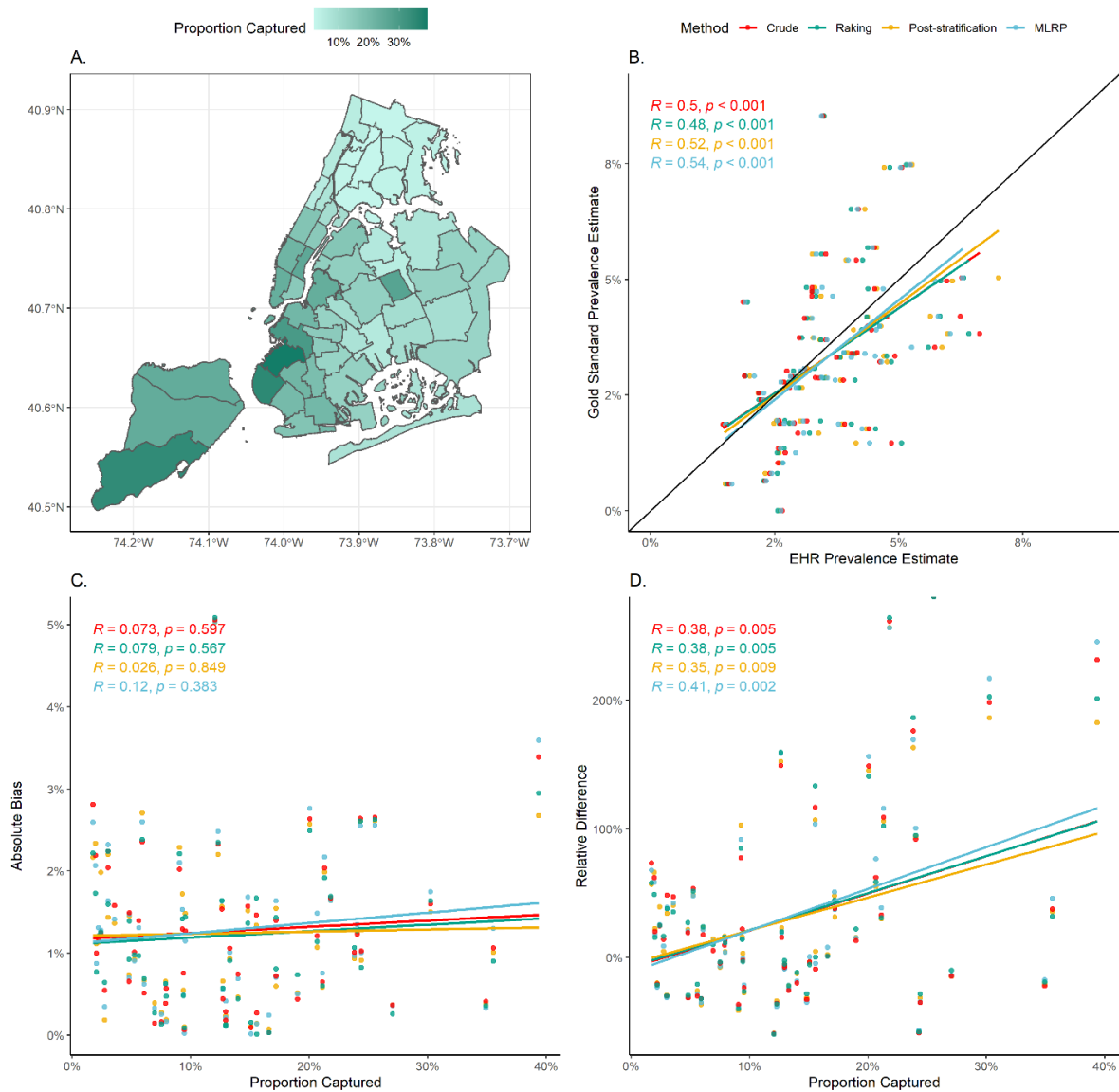


Figure 2: Characterization of the NYU Langone Patient Sample and Comparison of NYU EHR-Based to Gold Standard Diabetes Prevalence Estimates for Young Adults Aged 18-44 Years by New York City PUMA Neighborhood.²

² Panel 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. Panel B: Comparison of NYU EHR-based to gold standard diabetes prevalence estimates. Each point represents a PUMA neighborhood. EHR estimates are defined using NYU Langone Health 2019 data. The gold standard estimate for is defined using NYC CHS 2015-2020 data. Panel C: Comparison of absolute bias in NYU EHR-based prevalence estimates vs. proportion of the general population captured within the EHR sample. Absolute bias calculated as the absolute percentage point difference between the gold standard and EHR-based prevalence estimate for each NYC PUMA neighborhood. Panel D: Comparison of relative bias in NYU EHR-based prevalence estimates vs. proportion of the general

3.3.2 Simulation Analyses

Scenario 1

In scenario 1, crude diabetes prevalence within the simulated EHR sample had a negative relative bias due to the decreased odds of selection within demographic subgroups with higher odds of diabetes (Figure 3A). Relative biases increased slightly as the proportion of the total neighborhood population captured in the sample increased, ranging from an average of -27.4% (s.d.=14.0) when including all neighborhoods to an average of -39.4% (s.d.=21.8) when including neighborhoods within the top quintile of proportion of the population captured in the sample.

Raking, post-stratification, and the MLRP specification that did not include the neighborhood random effect had average relative biases below 3% across all sample definitions. Performance was more variable and coverage worsened (from ~40% to ~30%) as catchment areas became more restrictive. While the MLRP specification that did include the neighborhood random effect had better performance in the unrestricted sample (with a coverage of 92% and average relative bias of 1.1%, Appendix B Table 5), performance worsened, as measured by both average relative biases and coverage, as the catchment area definition became more restrictive (Figure 3A, Appendix B Table 5).

Scenario 2

In scenario 2, 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 3B). Adjustment methods partially accounted for this bias, however

population captured within the EHR sample. Relative bias calculated as the percent change between the gold standard and EHR-based prevalence estimate for each NYC PUMA neighborhood.

substantial residual biases remained, with coverage below 70% for all adjusted estimates (Appendix B 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%.

Scenario 3

In scenario 3, crude diabetes prevalence within the simulated EHR sample had an average relative bias ranging from -94% when those with diabetes had strong decreased odds of selection ($OR_{DM}=0.33$) to +148% when those with diabetes had strong increased odds of selection ($OR_{DM}=3.0$) (Figure 3C). 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 to crude estimates (Figure 3C).

Simulation results displayed similar patterns for neighborhood-level estimates (Appendix B Figure 1).

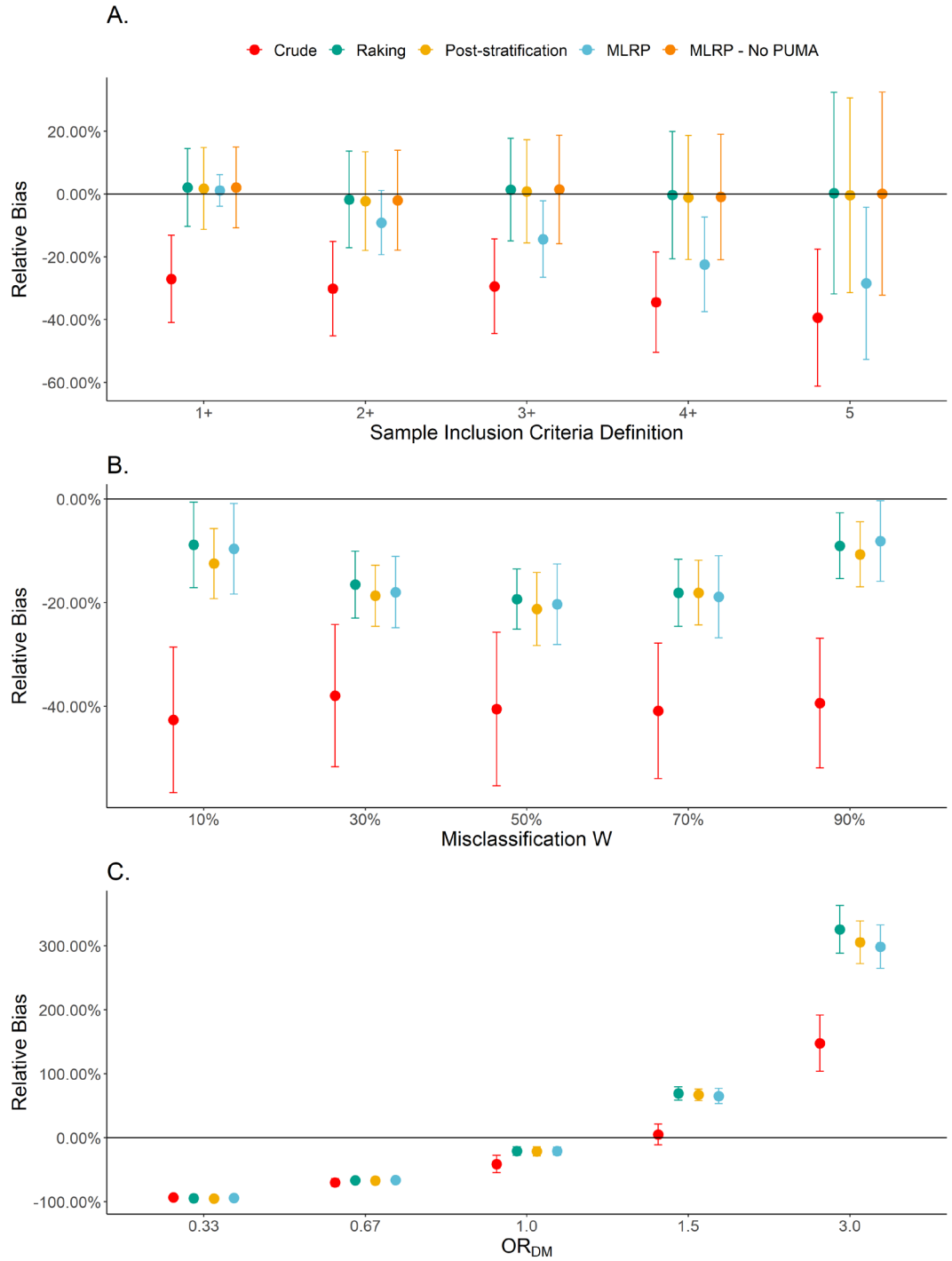


Figure 3: Mean Relative Bias in the EHR-Based Estimates vs. True Diabetes Prevalence by Simulation Scenario.³

³ Error bars represent standard deviation in mean relative bias across simulations. Panel A: Scenario 1 modified inclusion criteria definition based on quantiles of the proportion of the neighborhood general population captured in

3.4 Discussion

In this paper, bias adjustment methods were applied to real-world EHR data to explore whether valid diabetes prevalence estimates could be generated for young adults within NYC. Within the NYU Langone patient sample, crude diabetes prevalence was lower than the gold standard health survey-based estimate of diabetes prevalence for NYC young adults. All adjustment methods performed similarly and produced prevalence estimates that were systematically higher, yet statistically equivalent to gold standard estimates. This minor positive adjustment may have been partially driven by the relative down-weighting of individuals who were female, White, or residing in the top quartile of overrepresented neighborhoods, all of whom had lower diabetes prevalence within the EHR sample.

However, within neighborhood-level analyses, we observed that relative differences from gold standard estimates increased as proportion of the general population captured in the sample increased. This finding was counter-intuitive, as demographic representativeness increased with increased relative capture, and we would assume that representativeness on unmeasured factors would also increase, resulting in improved estimation in these neighborhoods. Differential patterns in selection into the EHR by neighborhood and health status could help explain this finding. For example, those with diabetes may have been more likely to be seen by the healthcare system within neighborhoods in close proximity to outpatient facilities, which could have resulted in unintended positive biases and contributed to a slight overestimation of diabetes prevalence within these neighborhoods and the overall citywide estimates. Further, including auxiliary PUMA-level SDOH and health outcomes had a minor impact on the results, increasing positive adjustments by less than one percentage point overall. Within this analysis, these

the sample; Panel B: Scenario 2 modified the level of misclassification of the auxiliary variable W compared to the unobserved variable U ; Panel C: Scenario 3 modified the association between diabetes and selection (OR_{DM}).

neighborhood-level covariates may not have been strong predictors of diabetes at this geographic scale.⁸⁶

Simulation analyses were then used to probe the potential for residual selection biases within EHR-derived estimates. Importantly, simulations demonstrated that MLRP models with a neighborhood random effect out-performed other adjustment procedures when samples did not exclude neighborhoods with lower capture of the general population. However, these models were highly sensitive and performance worsened as catchment areas became more restrictive. This illustrates how MLRP's partial pooling, or shrinkage towards the global mean, can improve precision when there are sparse strata but may begin to over-smooth estimates when there are large numbers of strata with no sampled individuals.⁸⁷ This simulation result conflicted with sensitivity analyses in the real-world analysis, where estimation to the NYC general population improved as catchment areas became more restrictive, perhaps reflecting a chance balancing of this shrinkage towards the global mean with unintended positive biases produced through the adjustment procedures. These patterns likely vary across diseases or age groups, and investigators should explore the impact of restricting EHR samples to health system catchment areas when making inferences to larger geographic areas of public health importance.

Simulations also demonstrated that adjustment methods were sensitive to selection biases by unobserved predictors of diabetes. While auxiliary information could help to account for these biases, the simulations demonstrate that residual biases will exist even when this auxiliary information is a strong proxy variable. This scenario is likely to occur in real-world analyses. Evidence supports that SDOH, such as income, language, culture, or education, are associated with diabetes and healthcare utilization.¹² However, these variables are notoriously difficult to measure using EHR data. Consistent with prior research, the use of Medicaid status and

neighborhood-level SDOH were imperfect proxies that may not have fully accounted for potential selection biases by these factors.^{88,89} Continued efforts to incorporate and utilize SDOH screening tools within EHRs, which can include detailed questions on food/housing insecurity, social isolation, stress, or other factors that affect use of the healthcare system, may improve estimation through these methods.⁹⁰ The use of PUMAs likely contributed to the poor predictive power of the neighborhood-level SDOH in this analysis. Neighborhood-level SDOH defined using smaller geographic areas, such as Zip Code Tabulation Areas, have been shown to improve estimation of diabetes prevalence through MLRP methods.⁷²

More significantly, simulations demonstrated that biases could be exacerbated through these methods when diabetes increased the odds of selection into the EHR sample. These SNAR scenarios are also plausible within the real-world, as individuals with chronic conditions may be more likely to receive regular care than individuals who are healthy.^{10,11,13} SNAR mechanisms could be further complicated by neighborhood. For example, patients residing in neighborhoods within close proximity, where capture of the general population within the EHR sample is high, may be more likely to use the health system for routine care, including diabetes management.¹⁰ The observed positive relative differences and positive trend between relative differences and proportion of the general population captured in the real-world sample could be partially attributed to such an SNAR mechanism. The tested adjustment methods are limited to including variables that are observed in both the sample and general population, and it is therefore difficult to account for SNAR mechanisms. Including neighborhood-level health outcomes in the MLRP models did not have a large impact on prevalence estimates, consistent with prior research using neighborhood hospitalization rates.⁷² As proposed in the missing data literature, additional

granular data on variables that are strongly correlated with diabetes (e.g., obesity) within the general population could improve these methods.⁷⁷

Based on the real-world and simulation analyses, we do not have confidence that these methods, as implemented, would consistently produce valid estimates of diabetes prevalence among young adults across jurisdictions or EHR data sources. The persistent positive relative differences compared to gold standard estimates supports the hypothesized presence of an SNAR mechanism, where those with diabetes are more likely to be users of healthcare systems, which could result in the overestimation of diabetes prevalence. 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. This potential could be realized by using population-representative clinical data sources (e.g., all-payer claims databases) to incorporate neighborhood-level healthcare utilization patterns or health outcomes at more granular geographic scales.

This study has a number of strengths, including simulation assumptions based on real-world analyses, use of large sample sizes, and external validation compared to gold standard surveillance estimates. However, there are a number of limitations to the gold standard data source. NYC CHS estimates are based on self-reported diabetes status, which could be under-reported from undiagnosed individuals who not in-care. The NYC CHS data were also pooled from 2015-2020 to produce reliable neighborhood-level prevalence estimates within the young adult age group. As diabetes prevalence has increased over time, this pooling could also contribute to lower prevalence within the gold standard estimates. This pooled time period also includes the start of the COVID-19 pandemic, during which time many ambulatory services were paused and access to telemedicine or eventual re-uptake of in-person care was differential across populations.^{91,92} Self-reported disease status from this year could be susceptible to

underestimation of disease burden from these disruptions to primary care services. These limitations could contribute to the positive relative differences observed within the real-world analyses.

Additionally, the computable phenotypes for diabetes status used in this analysis have not been validated by chart review within the NYU EHR data. While misclassification of diabetes status, including classifying those with prediabetes or undiagnosed diabetes as diabetic, could have contributed to the positive relative differences observed within the adjusted EHR-based estimates, studies elsewhere have shown that incorporating evidence across multiple sources (e.g., diagnoses and medications) helps to improve sensitivity, or the ability to detect true diabetes cases.^{1,44,54} Misclassification of disease status is also dependent on healthcare utilization patterns.¹¹ If patients within outlying neighborhoods that had a low proportion of the general population captured within the sample were also less frequent users of outpatient NYU clinics, they may have been less likely to receive A1C testing or diabetes diagnoses,¹⁰ which could also contribute to the observed positive trend between relative differences and relative capture. Additional work is needed to account for information biases within diabetes classification or within other EHR derived variables.⁹³ These results are also specific to diabetes within young adults, and selection biases likely differ across diseases or age groups.

Limitations were also present within the simulation analyses. While assumptions underlying the data generation process were based on real-world analyses, these likely represent a simplification of true selection processes. The simulations were time consuming and resource intensive, so scenarios that altered other pathways in the causal diagram were not explored.

3.5 Conclusion

While EHRs offer a rich source of clinical information, selection biases inherent in these data could limit their utility for population surveillance purposes and researchers should clearly communicate potential uncertainty or biases in their estimates. Statistical methods like MLRP could help to account for these biases. However, these methods 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. Additional research is needed to assess how more advanced methods, such as Bayesian modeling, could better handle SNAR mechanisms.

Chapter 4: Addressing Information Biases within Electronic Health Record Data for Examining Epidemiologic Associations with Diabetes Prevalence among Young Adults

4.1 Introduction

Understanding patterns and risk factors of chronic disease burden is a key function of public health practice. Recently, large, national research networks, including the Centers for Disease Control and Prevention (CDC)-funded Multi-state EHR-based Network for Disease Surveillance (MENDS) and Diabetes in Children and Young Adults (DiCAYA) networks, have been established to explore how electronic health records (EHRs) can be used for chronic disease surveillance.^{29,30} EHR data have numerous strengths that can be leveraged for this purpose. First, EHR data are available on large patient populations and offer near real-time information, allowing for improved precision and timeliness when estimating patterns or trends in disease burden.⁶⁷ Second, EHRs contain clinically-based diagnoses, lab results, and physical measurements.⁶⁸ Often, researchers use these clinical data to classify patients' disease status using rule-based computable phenotypes, or pre-specified logic-based criteria.⁹⁴ Using EHR-defined disease status would seemingly improve estimation of patterns or associations compared to some traditional surveillance systems (e.g., health surveys), which can have poor validity from self-reported disease status and limited reliability for small subgroups or geographies.⁶

However, while EHR data are clinically-based, the classification of disease status using these data is susceptible to information biases. Information biases represent systematic errors within estimates of epidemiologic associations arising from measurement error in key variables, like disease status. Within EHR data, a diagnostic suspicion bias could occur if certain patients

have increased screening for health outcomes (e.g., those who are obese may be more likely to be screened for diabetes with A1c testing),¹¹ and an informed presence bias could occur if, for example, patients who visit the health system more frequently are sicker or have more opportunities to receive a diagnosis.¹⁰ These biases could distort our understanding of patterns or risk factors for disease burden, such as by underestimating relative burden among those with decreased access or more fragmented care. Misclassification within EHR-based disease status can be measured through sensitivity and specificity. Sensitivity refers to the proportion of truly diseased individuals who are correctly identified as having the outcome, while specificity refers to the proportion of truly non-diseased individuals who are correctly identified as not having the outcome.⁹⁵ Misclassification increases as the sensitivity or specificity of the EHR-based computable phenotype decreases.

Methodological research has introduced a variety of statistical approaches and frameworks to help correct for misclassification. Two compelling frameworks for EHR-based research include treating misclassification as either a missing data problem or as a causal inference problem.⁹⁶ Under a missing data framework, the observed health outcome is assumed to have some level of misclassification and the true health outcome is treated as missing for some or all of the patients.^{96,97} Traditional missing data methods, such as regression calibration, multiple imputation, and inverse probability weighting (IPW), can then be used to correct for this misclassification.⁹⁶⁻⁹⁹ When applying this framework to EHR data, algorithms to define rare diseases generally have near-perfect specificity and misclassification can be hypothesized to largely occur from missing data among diseased individuals who are falsely classified as non-diseased.¹⁰⁰ For example, pertinent evidence of the disease may be absent from a single EHR due to patients receiving care across multiple distinct healthcare systems.¹¹ For non-rare outcomes,

researchers may rely on variables that are identifiably missing (e.g., absence of BMI measurement or lab result) or on internal or external validation datasets that allow for the estimation of sensitivity and specificity of the computable phenotype.^{96,98}

Under the causal inference framework, misclassification in disease status can be conceptualized using directed acyclic graphs (DAGs).¹⁰¹ Non-differential misclassification in the exposure (E) or outcome (O) is characterized by independence between errors that arise in the observed exposure (E*) and outcome (O*), visually displayed as no unblocked backdoor paths between E and O (Figure 4A). Within this scenario, non-differential misclassification would generally result in biases towards the null, or an underestimation of the association with the health outcome.⁹⁶ More problematically, differential misclassification is characterized by dependence between these errors, or unblocked backdoor paths between E and O (Figure 4B-C).^{96,101} Differential misclassification can result in biases towards or away from the null, meaning that this has the potential to induce an association that otherwise does not exist.⁹⁶

Based on hypothesized DAGs, researchers can then use traditional epidemiologic methods to control for variables (U) that act as confounders of E* and O*. Importantly, controlling for common effects of E* and O* can inadvertently produce a collider or M bias. Further, a Berkson's bias can be produced when the sample is selected or restricted based on a common effect of E* and O*.¹³ For example, researchers often hypothesize that the number of healthcare encounters will affect misclassification of EHR-defined disease status through an informed presence bias, with EHR samples further restricted to those with at least one encounter.¹⁰ Prior studies have demonstrated that conditioning on the number of encounters could reduce confounding from differential misclassification but has the potential to induce a

smaller Berkson's or M bias if the number of encounters is a common effect of the exposure and outcome (e.g., Figure 4C), particularly when computable phenotypes are highly sensitive.¹³

In this paper, we explored the impact of information biases within EHR data on epidemiologic research questions related to diabetes among young adults aged 18-44 years. Diabetes is serious, chronic condition that is increasing in both prevalence and incidence^{20,102,103} but is still relatively rare within this age group, affecting an estimated 3% of those aged 18-44 years.⁶⁵ Our research aims were (1) to characterize age-and-sex adjusted odds of diabetes by race/ethnicity, an established social determinant of health (SDOH) associated with this condition,^{65,104,105} and (2) to test for an association between asthma and diabetes, two chronic conditions that have previously been observed to be associated but have an uncertain causal relationship.¹⁰⁶ We assessed the ability to correct for misclassification in diabetes status using the missing data and causal inference frameworks and compared naïve and corrected associations observed within EHR data to associations estimated from national health surveys. The overarching goal of this paper was to inform the broader discussion on how to address misclassification of disease outcomes within EHR data in order to provide valid estimates of diabetes prevalence for public health surveillance.

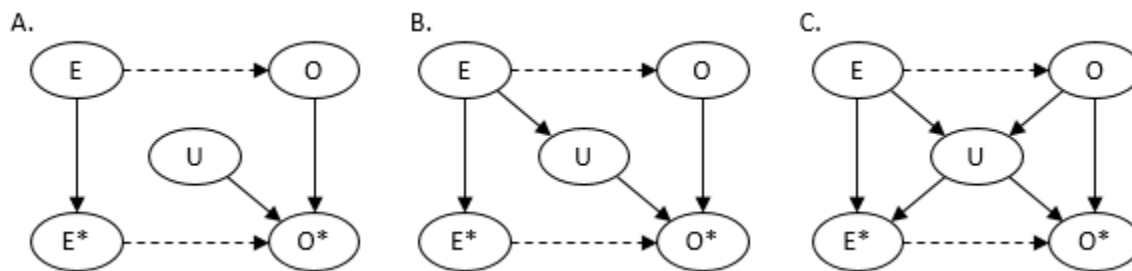


Figure 4: Directed Acyclic Graphs (DAGS) for Measurement Error in Exposure and Outcome Variables⁴

⁴Panel A: Non-differential misclassification of exposure and outcome. Panel B: Differential misclassification where E affects measurement of O through U (e.g., a diagnostic suspicion bias where those who are obese (E) are more

4.2 Methods

4.2.1 Data Sources

The EHR sample was composed of patients from NYU Langone Health, a large academic medical center with primary service areas in the New York City (NYC) boroughs of Manhattan, Brooklyn, and Queens. EHR data were pulled from the Epic Clarity database for all NYC-resident patients aged 18-44 years who had an inpatient or outpatient encounter from 2017-2019. Patient addresses were geocoded using the NYC Department of City Planning's NYCgbat Geosupport Desktop Edition software through the "rGBAT16AB" R package.¹⁰⁷ Demographic variables of age, sex (male or female), race/ethnicity (White, Black, Latino, Asian, and Other), insurance status (Medicaid vs. non-Medicaid), and neighborhood poverty level (<10%, 10-<20%, 20-<30%, and \geq 30% living in poverty within resident census tract) were defined for each patient. Race/ethnicity was imputed for those with unknown race/ethnicity (19%) using the Bayesian Improved Surname Geocoding (BISG) methods through the "wru" R package.⁷⁸ Neighborhood poverty level was assigned using zip code tabulation area (ZCTA) poverty group when census tract level data was unavailable (1%). Those with an unknown/other age or sex (<1%) were excluded from all analyses.

Healthcare utilization variables of total encounters, duration within the EHR system, presence of at least one routine health exam (ICD-10-CM: Z00.*), presence of at least one diabetes-related lab (fasting glucose, random glucose, or A1C), and presence of at least one encounter with an endocrinology review of systems were also defined. Endocrinology review of

likely to be screened for diabetes with A1c testing (U)). Panel C: Differential misclassification where U affects measurement of both E and O and is a common effect of E and O (e.g., an informed presence bias where diseases E and O increase individuals' number of healthcare encounters (U), which in turn increases the odds of receiving a diagnosis for E or O).

systems were identified using keyword text searches of History and Physical Examination (H&P) notes and progress notes (Appendix C).

Patients were classified as having prevalent obesity, asthma, and diabetes if they had evidence supporting these chronic conditions, using all historical EHR data through 2019, and were classified as not having each respective health outcome without such evidence. In alignment with diabetes definitions from national health surveys, EHR-defined diabetes included all diabetes types (Type 1, Type 2, and other). Evidence of diabetes included at least two encounter diagnoses for diabetes (ICD-10-CM: E08.*, E09.*, E10.*, E11.*, and E13.*), one encounter diagnosis and at least two elevated A1C lab results $\geq 6.5\%$, or at least one anti-diabetes prescription medication (not including metformin or acarbose).⁷⁹ Evidence of asthma was defined as at least two encounter diagnoses for asthma (ICD-10-CM: J45*–J46*) or at least two prescriptions for asthma-related medications.⁷² To maintain consistency across chronic disease classification methods, evidence of obesity was defined as a most recent BMI of at least 30 kg/m², with no naïve corrections for those who were missing BMI, height, or weight measurements (19%).

For comparison to traditional surveillance systems, samples were obtained from two publicly-available national health surveys, 2019 Behavioral Risk Factor Surveillance System (BRFSS) and pooled 2013-March 2020 National Health and Nutrition Examination Survey (NHANES). BRFSS is a cross-sectional telephone survey conducted by the CDC annually on a sample of over 400,000 United States (US) adults.¹⁰⁸ Within BRFSS data, diabetes and asthma were defined by self-reported prior diagnosis from a medical provider, and obesity was defined by a BMI of at least 30 kg/m² based on self-reported height and weight. Demographic variables of 5-year age group, sex (male or female), race/ethnicity (White, Black, Latino, Asian, and

Other), insurance status (uninsured vs. insured), and income level (<\$50K, \$50-<\$75K, ≥\$75K) were defined for each respondent. To reduce the effects of undiagnosed diabetes on misclassification of self-reported diabetes status, the BRFSS survey data were subset to those aged 18-44 years who were in-care, as defined as those who reported having a personal healthcare provider. Sensitivity analyses varied this definition to include those who were not in-care.

NHANES is a cross-sectional survey involving interviews and physical examinations that is conducted by the CDC annually on a sample of approximately 5,000 US children and adults.¹⁰⁹ Within NHANES data, diabetes was defined by self-reported prior diagnosis or elevated lab results (A1c ≥6.5% or fasting glucose ≥126 mm/Hg).^{65,110} Sensitivity analyses varied this definition to be based solely on self-reported prior diagnosis. Asthma was defined by self-reported prior diagnosis from a medical provider and obesity was defined by a measured BMI of at least 30 kg/m². Demographic variables of age, sex (male or female), race/ethnicity (White, Black, Latino, Asian, and Other), insurance status (Medicaid vs. non-Medicaid), and income level (≤130%, 130-350%, >350% of the federal poverty level) were defined for each respondent. NHANES data were subset to those aged 18-44 years. As lab results were used to supplement self-reported diabetes status, no exclusions were made based on in-care status.

4.2.2 Statistical Analyses

We estimated odds ratios (OR) for diabetes by race/ethnicity and asthma status under four frameworks that we describe herein. First, naïve models were estimated by fitting a logistic regression model for observed diabetes status (DM*) on the full patient sample. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex,

race/ethnicity, Medicaid insurance status, obesity, and neighborhood poverty level, as informed by the literature.

Second, complete case models were estimated among the subset of patients who we hypothesized to have complete data, defined as those with at least one encounter with an endocrinology review of systems or those who were classified as diabetic through the above definition. Since diabetes is a rare disease within the young adult population, we assumed that specificity of the classification was near-perfect and all patients who were classified as diabetic were complete cases. Sensitivity analyses tested this assumption and varied the definition of complete case to incorporate information related to the other health outcomes (e.g., having a BMI measurement or respiratory review of systems) (Appendix C). Complete case models were estimated by fitting a logistic regression model for DM* using the same covariates as the naïve models.

Third, under the missing data framework, we hypothesized that missing health outcomes would be predicted by demographics (e.g., differential screening by race/ethnicity), healthcare utilization (e.g., informed presence bias), and neighborhood (e.g., degree of continuity of care within the health system by catchment area). We estimated the probability of being a complete case using a multilevel logistic regression model including all demographic and healthcare utilization variables and a random intercept for neighborhood defined by Public Use Microdata Areas. Stabilized IPW weights were then calculated as the inverse of the predicted probability of being a complete case multiplied by the overall probability of being a complete case.⁹⁸ The final missing data models were estimated by fitting a logistic regression model for DM* on the subset of patients defined as having complete data, weighted for the stabilized IPW weights and using the same covariates as the naïve models.

Fourth, under the causal framework, we hypothesized that total encounters would be associated with misclassification of health outcomes through an informed presence bias and would be an effect of health outcomes, consistent with prior research.^{10,13} Hypothesized DAGs underlying the causal framework are included in Appendix C. Causal framework models were estimated by fitting a logistic regression model for DM* using the full patient sample, controlling for total encounters and the covariates included in the naïve model.

EHR-derived ORs were compared to ORs from the two health surveys, which were obtained by fitting logistic regression models for diabetes accounting for the complex sample designs. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex, race/ethnicity, obesity, insurance status, and income level. Relative differences between EHR-based and health survey-based ORs were calculated as percent differences.

4.3 Results

The EHR sample was composed of 454,612 patients seen within the NYU Langone sample from 2017-2019, with mean age of 32 years (Table 4). A total of 37.8% of patients were male and 22% had Medicaid insurance. The largest racial/ethnic group within the sample was White, with 42% having a White race/ethnicity recorded within the EHR and 52% classified as White through BISG imputation. Approximately one-quarter of patients had a routine medical exam and one-half had a DM-related lab. Within the full sample, 3.1% of patients were classified as having diabetes, 17.5% were classified as being obese, and 4.2% were classified as having asthma.

A total of 40% of the patient population were classified as complete cases (Table 4). Patients who were classified complete cases had greater healthcare utilization, measured by a higher average number of total encounters, greater duration within the NYU system, and greater

proportion having at least one BMI, routine health exam, or diabetes-related lab. Compared to non-complete cases, a greater proportion also had a known race/ethnicity (87.4%) or were classified as obese (22.5%) or asthmatic (7.2%).

Table 4. Descriptive Summary of NYU Patient Population by Complete Case Status.

	Total Sample	Non-Complete Case	Complete Case^a
Total	454,612	273,576 (60.4%)	181,036 (39.8%)
Age (mean (sd))	32.13 (7.11)	32.02 (7.13)	32.31 (7.07)
Sex (Male)	171968 (37.8)	100964 (36.9)	71004 (39.2)
Medicaid Insurance (Yes)	100979 (22.2)	59001 (21.6)	41978 (23.2)
Raw Race/Ethnicity			
White	190225 (41.8)	111123 (40.6)	79102 (43.7)
Black	45509 (10.0)	24442 (8.9)	21067 (11.6)
Latino	62989 (13.9)	32157 (11.8)	30832 (17.0)
Asian/PI	35262 (7.8)	20947 (7.7)	14315 (7.9)
Other	32525 (7.2)	19669 (7.2)	12856 (7.1)
Missing	88102 (19.4)	65238 (23.8)	22864 (12.6)
Imputed Race/Ethnicity			
White	237057 (52.1)	144783 (52.9)	92274 (51.0)
Black	57709 (12.7)	33439 (12.2)	24270 (13.4)
Latino	86679 (19.1)	49131 (18.0)	37548 (20.7)
Asian/PI	49170 (10.8)	31288 (11.4)	17882 (9.9)
Other	23997 (5.3)	14935 (5.5)	9062 (5.0)
Any Recorded BMI (Yes)	367903 (80.9)	190916 (69.8)	176987 (97.8)
All Encounters (mean (sd))	15.23 (23.51)	9.41 (13.15)	24.03 (31.60)
Duration (mean (sd))	1.84 (1.93)	1.53 (1.83)	2.31 (1.98)
≥ 1 Routine Medical Exam (Yes)*	115249 (25.4)	32786 (12.0)	82463 (45.6)
≥ 1 DM-related Lab (Yes)^b	205408 (45.2)	80728 (29.5)	124680 (68.9)
PUMA Coverage			
< 10 %	79563 (17.5)	48320 (17.7)	31243 (17.3)
10-<20%	143907 (31.7)	82148 (30.0)	61759 (34.1)
20-<30%	163076 (35.9)	101293 (37.0)	61783 (34.1)
30-<40%	68066 (15.0)	41815 (15.3)	26251 (14.5)
Asthma (Yes)	19240 (4.2)	6167 (2.3)	13073 (7.2)
Obese (Yes)	79580 (17.5)	38819 (14.2)	40761 (22.5)

^a Complete cases defined as those with at least one encounter with an endocrinology review of systems or those who were classified as diabetic through the computable phenotype of having at least two encounter diagnoses for diabetes, one encounter diagnosis and at least two elevated A1C lab results $\geq 6.5\%$, or at least one anti-diabetes prescription medication.

^b Including all A1c, random blood glucose, and fasting blood glucose lab results.

*Defined within the years of 2017-2019.

Within the full (naïve) EHR sample, those who were classified as non-diabetic consistently had lower healthcare utilization than those who were classified as diabetic (Table 5).

This pattern was disrupted in the complete case sample, where a greater proportion of non-diabetic patients had at least one routine medical exam (47% versus 32% of diabetic patients) and almost all patients had a recorded BMI regardless of diabetes status. Within both the naïve and complete case samples, a lower proportion of those classified as non-diabetic were of Black or Latino race/ethnicity and were classified as having asthma or obesity.

Table 5: Descriptive Summary of NYU Patient Population by Diabetes Status.

	Naive Diabetes Status		Complete Case Diabetes Status ^a	
	Non-Diabetic	Diabetic	Non-Diabetic	Diabetic
Total	440568 (96.9)	14044 (3.1)	166992 (92.2)	14044 (7.8)
Age (30-44 years)	273216 (62.0)	10843 (77.2)	104005 (62.3)	10843 (77.2)
Sex (Male)	166204 (37.7)	5764 (41.0)	65240 (39.1)	5764 (41.0)
Medicaid Insurance (Yes)	96760 (22.0)	4219 (30.0)	37759 (22.6)	4219 (30.0)
Raw Race/Ethnicity				
White	185119 (42.0)	5106 (36.4)	73996 (44.3)	5106 (36.4)
Black	43289 (9.8)	2220 (15.8)	18847 (11.3)	2220 (15.8)
Latino	59636 (13.5)	3353 (23.9)	27479 (16.5)	3353 (23.9)
Asian/PI	34174 (7.8)	1088 (7.7)	13227 (7.9)	1088 (7.7)
Other	31378 (7.1)	1147 (8.2)	11709 (7.0)	1147 (8.2)
Missing	86972 (19.7)	1130 (8.0)	21734 (13.0)	1130 (8.0)
Imputed Race/Ethnicity				
White	231412 (52.5)	5645 (40.2)	86629 (51.9)	5645 (40.2)
Black	55268 (12.5)	2441 (17.4)	21829 (13.1)	2441 (17.4)
Latino	82827 (18.8)	3852 (27.4)	33696 (20.2)	3852 (27.4)
Asian/PI	47887 (10.9)	1283 (9.1)	16599 (9.9)	1283 (9.1)
Other	23174 (5.3)	823 (5.9)	8239 (4.9)	823 (5.9)
Any Recorded BMI (Yes)	354043 (80.4)	13860 (98.7)	163127 (97.7)	13860 (98.7)
Obese (Yes)	73412 (16.7)	6168 (43.9)	34593 (20.7)	6168 (43.9)
All Encounters (mean (sd))	14.29 (21.16)	44.90 (54.27)	22.27 (28.19)	44.90 (54.27)
Duration (mean (sd))	1.81 (1.92)	2.87 (2.06)	2.26 (1.97)	2.87 (2.06)
≥1 Routine Medical Exam (Yes)*	110713 (25.1)	4536 (32.3)	77927 (46.7)	4536 (32.3)
≥ 1 DM-related Lab (Yes)^b	193480 (43.9)	11928 (84.9)	112752 (67.5)	11928 (84.9)
PUMA Coverage				
< 10 %	76463 (17.4)	3100 (22.1)	28143 (16.9)	3100 (22.1)
10-<20%	139333 (31.6)	4574 (32.6)	57185 (34.2)	4574 (32.6)
20-<30%	158890 (36.1)	4186 (29.8)	57597 (34.5)	4186 (29.8)
30-<40%	65882 (15.0)	2184 (15.6)	24067 (14.4)	2184 (15.6)
Asthma (Yes)	17339 (3.9)	1901 (13.5)	11172 (6.7)	1901 (13.5)

^a Complete cases defined as those with at least one encounter with an endocrinology review of systems or those who were classified as diabetic through the computable phenotype of having at least two encounter diagnoses for diabetes, one encounter diagnosis and at least two elevated A1C lab results $\geq 6.5\%$, or at least one anti-diabetes prescription medication.

^b Including all A1c, random blood glucose, and fasting blood glucose lab results.

**Defined within the years of 2017-2019.*

In both the survey and EHR-based analyses, respondents who were Black or Latino had significantly higher odds of diabetes compared to White respondents, controlling for age and sex (Figure 5). The naive EHR-based OR estimate for Latino patients was 26% higher than the BRFSS point estimate ($OR_{\text{Naive}}=1.93$, 95% CI: 1.85-2.01 vs. $OR_{\text{BRFSS}}=1.53$, 95% CI: 1.32-1.77). All correction methods reduced this association, with point estimates falling within the 95% confidence intervals and relative differences below 15% compared to both health survey estimates. BRFSS and NHANES respondents who were Asian did not have significantly higher odds of diabetes compared to White respondents, and confidence intervals were wide due to small sample sizes of this subgroup. Within the EHR analyses, patients who were Asian had a significant 11-26% increased odds of diabetes.

In the BRFSS and NHANES analyses, having asthma was associated with an approximate 20-40% increased odds of diabetes after controlling for demographics and obesity ($OR_{\text{BRFSS}} = 1.23$, 95% CI: 1.09-1.40; $OR_{\text{NHANES}} = 1.38$, 95% CI: 1.01-1.91). In the naive EHR analysis, asthma was strongly associated with diabetes, with those with asthma estimated to have three times the odds of diabetes as those without asthma after controlling for demographics and obesity (95% CI: 2.86-3.18). This association was reduced in the complete case sample and the IPW-weighted complete case sample, with corrected OR estimates falling within the 95% CI of the NHANES estimate and having an approximate 30-50% relative difference from the health survey point estimates. The association between asthma and diabetes was further reduced in the causal model ($OR = 1.42$, 95% CI: 1.34-1.51), with a 3% relative difference from the NHANES point estimate and a 15% relative difference from the BRFSS point estimate (Figure 5). Sensitivity analyses varying the definition for complete cases, varying the BRFSS inclusion

criteria, or varying the NHANES diabetes definition produced similar patterns in these results (Appendix C).

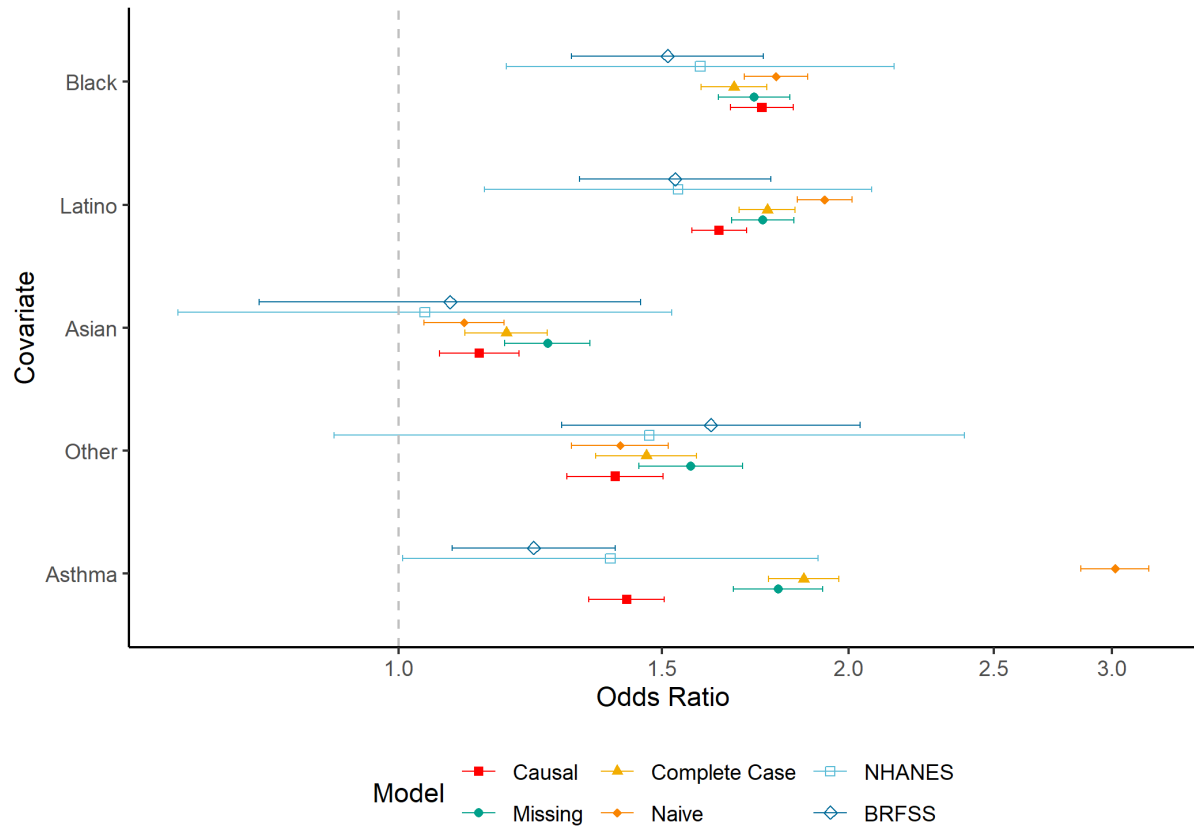


Figure 5: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, EHR-Based Estimates vs. Health Survey Estimates⁵

⁵ Associations with self-reported diabetes diagnosis observed among Behavioral Risk Factor Surveillance System Survey 2019 respondents aged 18-44 years who report having a routine checkup within the past two years. ORs for race/ethnicity were adjusted for age group and sex, and ORs for asthma were adjusted for age group, sex, race/ethnicity, obesity, insurance status, and income level. Associations with self-reported diabetes diagnosis or undiagnosed diabetes (A1C \geq 6.5% or fasting glucose \geq 126mm/Hg) observed among National Health and Nutrition Examination Survey 2013-March 2020 respondents aged 18-44 years. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex, race/ethnicity, obesity, insurance status, and income level. Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2019. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex, race/ethnicity, obesity, insurance status, and neighborhood poverty level. Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2019 with complete records, as defined as those with a review of systems for endocrinology. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex, race/ethnicity, obesity, insurance status, and neighborhood poverty level. Associations with EHR-

4.4 Discussion

Although EHRs offer clinically-based measurements that may improve disease classification compared to self-report, these data have the potential for information biases arising from situations like over-screening within select populations or an informed presence bias. Our analysis explored the potential impact of information biases on observed associations of diabetes risk factors within an EHR sample of young adults. We found that those who were observed to have diabetes had greater healthcare utilization than those who were classified as non-diabetic, which could contribute to such an information bias. For example, those with a greater number of healthcare encounters may have been more likely to have documentation an underlying diabetes or asthma diagnosis. Imposing complete case criteria to subset to those for whom we had greater confidence in an accurate diabetes classification helped to reduce disparities in healthcare utilization patterns by observed diabetes status. Additional methods based on missing data and causal frameworks were also tested to help correct for these biases.

Within naïve analyses, the observed age-and-sex adjusted odd ratios for diabetes among Latino patients appeared slightly inflated compared to health survey estimates. Subsetting analyses to complete cases, using IPW, or controlling for the number of healthcare encounters produced odds ratios that were closer to health survey estimates. Prior research has demonstrated that Latino and Black individuals may have increased screening for diabetes while Asian individuals may have decreased screening compared to White.¹¹¹ These disparities in screening

defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2019 with complete records, as defined as those with a review of systems for endocrinology, using missing data framework weighting the complete case subset with stabilized IPW. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex, race/ethnicity, obesity, insurance status, and neighborhood poverty level. Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2019, using the causal framework controlling for total number of encounters. ORs for race/ethnicity were adjusted for age and sex, and ORs for asthma were adjusted for age, sex, race/ethnicity, obesity, insurance status, and neighborhood poverty level.

practices could explain the observed patterns within the EHR estimates. Increased likelihood of screening would produce a positive bias while decreased likelihood of screening would produce a negative bias in naïve EHR associations since those who are not screened may have undiagnosed diabetes. The tested methods may have helped correct for this bias by controlling or restricting based on factors associated with likelihood of diabetes screening, resulting in decreases to the Black and Latino ORs and increases to the Asian OR relative to naïve estimates. Overall, determining the accuracy of the EHR-based estimates was challenging due to the wide confidence intervals for survey-based estimates, but EHR-based associations had greatly improved precision compared to gold standard due to the diversity and larger sample size of these data.

Consistent with prior research, the naïve association between two EHR-observed conditions, asthma and diabetes, was substantially positively biased relative to health survey estimates and prior studies from the literature.^{10,11,13,106} All correction methods greatly reduced the estimated association between these two chronic conditions, with the causal framework having the largest impact on this estimate. Since the number of healthcare encounters may be a common effect of these chronic conditions, it is possible that controlling for this variable in this correction method induced a small M-bias, producing an estimate that was lower than the complete case or missing framework estimates.¹³ However, all corrected EHR estimates were still higher than the health survey point estimates, suggesting that the NYU patient population may not be generalizable or that there are residual biases in these estimates. For example, those interacting with the NYU hospital system may be sicker and more likely to have multiple chronic conditions than those who receive care at independent primary care practices or those who are

not in care. This selection bias was not addressed in this work and is an important avenue for future research.

This study applied multiple bias correction frameworks to a large, diverse patient population and these findings can inform broader discussions on addressing misclassification of disease outcomes within epidemiologic studies using EHR data. However, there are limitations to these analyses. Methods focused on addressing misclassification of health outcomes, but there is potential for misclassification within other covariates. The hypothesized DAG (Appendix C) likely represents a simplified depiction of information biases within these data. In particular, a large proportion of patients had an unknown race/ethnicity, and the BISG imputation methods used may have differential performance by race/ethnicity or marital status.¹¹² Internal or external validation samples were also not available to inform computable phenotype sensitivity or specificity for the correction methods,⁹⁶ and the complete case and missing data frameworks relied on the assumption that the specificity of the computable phenotype was near perfect. Sensitivity analyses were used to test these assumptions, but it is possible that imposing complete case determinations generated a selection bias for sicker individuals who were more frequent users of the healthcare system,¹¹ which could explain why these methods found higher odds of diabetes among those with asthma compared to the causal framework methods. That said, internal and external validation samples are often costly or time-intensive to obtain, so these methods offer an imperfect, yet feasible, solution within resource-constrained environments.

Additionally, comparisons were made to estimates from two health surveys, which have distinct biases that were not addressed in this analysis. The BRFSS is limited to self-reported health outcomes, which can be prone to misclassification.^{4,6} Physical measurements from NHANES may improve classification of diabetes status among those who are unaware or

undiagnosed.⁶⁵ However, the smaller sample size of this survey requires multi-year pooled analyses, which can be biased from changes in screening or diagnostic criteria over time. For example, in 2015, the American Diabetes Association lowered the recommended BMI screening threshold for Asian Americans to better account for differential risk of diabetes at equivalent BMI levels, which could change the burden of undetected diabetes within this subgroup across time.¹¹³ While the complementary strengths of these two data sources may help to remedy these unaddressed biases, both data sources reflect national data, which may not be transportable to this NYC patient population. Although local versions of these health surveys are available, sample sizes were too small to produce reliable associations. Some covariate definitions also varied across data sources. Importantly, individual-level income was available within the survey data but was unavailable in the EHR data. The use of neighborhood-level poverty likely resulted in residual confounding in all EHR-based ORs for asthma, which may have contributed to the positive relative differences compared to survey-based estimates.

Achieving health equity is a core mission of public health practice. This study demonstrated the potential advantage of EHR data to better understand differential patterns and estimate absolute chronic disease burden among small racial/ethnic groups within local jurisdictions. These data also offer the opportunity to better understand implicit biases in how providers interact with patients or interface with EHR systems. Implicit biases, including stigmatizing medical notes or differential care received,^{111,114} affect both the quality of data within EHRs and, more importantly, the health and wellbeing of populations served by these systems. Future research is needed to further explore and remedy these biases. Additionally, this study assessed associations between prevalent asthma and diabetes due to uncertain temporality

within the survey data. However, the longitudinal nature of EHR data could be leveraged to estimate average causal effects between these chronic conditions.

4.2 Conclusion

EHRs offer a compelling data source for public health research, however, misclassification of disease status has the potential to bias results of these studies. Methods to treat misclassification as a missing data problem, to control for factors that affect misclassification using a causal framework, or to simply subset analyses to patients with complete data should be considered to help produce valid estimates of risk factor associations. More largely, these methods may help mitigate biases in estimates of absolute disease prevalence by demographics or other subpopulations of interest when using EHR data for chronic disease surveillance. Additional methods are needed to account for selection biases within EHR data to inform the transportability or generalizability of results.

Chapter 5: Conclusion

Using electronic health record (EHR) data to supplement existing public health surveillance systems could improve the timeliness, granularity, and validity of estimates of patterns and trends in chronic disease burden. However, these data are susceptible to selection biases from using a convenience sample of in-care patients and to information biases from using health outcomes measured by data systems designed for documentation of clinical care rather than research purposes. This integrated learning experience (ILE) explored such biases when using EHR data to understand prevalence and risk factors of diabetes among young adults aged 18-44 years. Diabetes is a rare yet serious chronic disease that is fast-evolving within this age group and could benefit from innovations to traditional surveillance systems.

The results of this ILE demonstrated the importance of applying an epidemiologic perspective when using EHR data for diabetes surveillance. First, the literature review illustrated how diabetes computable phenotypes, or case definitions, should be designed to reflect the analytic purpose, data sources, and target population of inference. Highly specific definitions were observed to have lower sensitivity compared to definitions that were more flexible or inclusive of different source of information from the EHR. Such overly restrictive case definitions have significant implications when EHR data are used to further health equity, a key purpose of public health surveillance. For example, a clinically consistent definition of diabetes requires two positive lab results. However, this definition would differentially misclassify individuals with decreased access to care who only have a single lab conducted as non-diabetic, leading to systematic underestimation of diabetes burden within populations with limited access to care.

Second, an exploration of selection biases from a convenience sample of patients from an academic medical center in New York City and from data simulations demonstrated the importance of understanding the underlying factors that affect which patients are included within an EHR sample. When common statistical methods that adjust for variables that are measured in both the EHR and general population were applied to these data, neighborhood-level proxies of social determinants or health status could not fully account for corresponding individual-level variables that affected selection processes. Hypothesized relationships between these factors can help investigators predict the magnitude and direction of residual biases, which is essential when communicating the uncertainty in estimated prevalence estimates, particularly when uncertainty may be differential across subgroups.

Third, an exploration of information biases from health outcomes classified using EHR data from the same academic medical center in New York City illustrated how missing data and causal inference frameworks can be applied to mitigate biases in the measurement of health outcomes with these data. Using inverse probability weighting or controlling for factors that were hypothesized to be associated with misclassification of health outcomes greatly reduced the strength of the association between asthma and diabetes status compared to naïve associations observed within the patient sample, in alignment with hypothesized directed acyclic graphs and with associations observed with national health survey data and in the literature. Without these adjustments, observed associations with EHR-defined diabetes could reflect differences in healthcare utilization and provide a biased understanding of patterns and risk factors of disease burden.

Limitations to this work include the lack of true gold standards for external validation. EHR-based estimates of prevalence and measures of association were compared to estimates

from local and national health surveys. These health surveys have known limitations, including poor validity from self-reported disease status, selection biases from low participation rates, and reduced reliability within underrepresented groups. These limitations likely contributed to the observed differences between EHR and survey-based estimates, which made it challenging to evaluate the extent of biases within the EHR-based estimates. In addition, this work was not an exhaustive evaluation of all biases that may be present within these data. While these findings may apply more broadly to the surveillance of chronic conditions, further research is needed to understand how biases may operate differently across diseases, data sources, or jurisdictions.

Innovations in chronic disease surveillance systems are needed to provide more timely and granular estimates of disease burden to better target public health resources to at-risk populations and to achieve health equity. As seen through this work, EHR data offer the opportunity to provide precise estimates of diabetes burden and risk factor associations for small geographic areas or demographic groups. However, care is needed when analyzing and interpreting these data to address inherent selection and information biases. Additional work is needed to improve the collection of social determinants of health, such as income level or education, within EHR data to better address these biases. Moreover, while these data offer great promise as a supplement to existing surveillance systems, they do not replace the need for health surveys or other forms of data collection that gather information on those who are not in-care. Collaborations between academic medical centers and public health agencies should be established to allow for the practical application of these data.

References

1. McVeigh KH, Lurie-Moroni E, Chan PY, et al. Generalizability of Indicators from the New York City Macroscopic Electronic Health Record Surveillance System to Systems Based on Other EHR Platforms. *EGEMS (Wash DC)*. 2017;5(1):25.
2. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. *Ann Epidemiol*. 2007;17(9):643-653.
3. Laflamme DM, Vanderslice JA. Using the Behavioral Risk Factor Surveillance System (BRFSS) for exposure tracking: experiences from Washington State. *Environ Health Perspect*. 2004;112(14):1428-1433.
4. Bowlin SJ, Morrill BD, Nafziger AN, Jenkins PL, Lewis C, Pearson TA. Validity of cardiovascular disease risk factors assessed by telephone survey: the Behavioral Risk Factor Survey. *Journal of clinical epidemiology*. 1993;46(6):561-571.
5. Gillum R, Sempos CT. Ethnic variation in validity of classification of overweight and obesity using self-reported weight and height in American women and men: the Third National Health and Nutrition Examination Survey. *Nutr J*. 2005;4(1):27.
6. Merrill RM, Richardson JS. Validity of self-reported height, weight, and body mass index: findings from the National Health and Nutrition Examination Survey, 2001-2006. *Preventing chronic disease*. 2009;6(4):A121.
7. Sidebottom AC, Johnson PJ, VanWormer JJ, Sillah A, Winden TJ, Boucher JL. Exploring electronic health records as a population health surveillance tool of cardiovascular disease risk factors. *Population health management*. 2015;18(2):79-85.
8. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health*. 2015;36:345-359.
9. Drawz PE, Archdeacon P, McDonald CJ, et al. CKD as a Model for Improving Chronic Disease Care through Electronic Health Records. *Clin J Am Soc Nephrol*. 2015;10(8):1488-1499.
10. 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(1):22-22.
11. 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. *Current epidemiology reports*. 2017;4(4):346-352.
12. 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(1):E28-32.

13. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *Am J Epidemiol*. 2016;184(11):847-855.
14. Hubbard RA, Huang J, Harton J, et al. A Bayesian latent class approach for EHR-based phenotyping. *Statistics in medicine*. 2019;38(1):74-87.
15. Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. *Journal of the American Medical Informatics Association*. 2019;26(8-9):730-736.
16. *Diabetes Public Health: From Data to Policy*. Oxford University Press; 2010.
17. Riddle MC, Herman WH. The Cost of Diabetes Care—An Elephant in the Room. *Diabetes care*. 2018;41(5):929-932.
18. Bullard KM, Cowie CC, Lessem SE, et al. Prevalence of Diagnosed Diabetes in Adults by Diabetes Type - United States, 2016. *MMWR Morbidity and mortality weekly report*. 2018;67(12):359-361.
19. Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. *JAMA*. 2014;312(12):1218-1226.
20. Gregg EW, Li Y, Wang J, et al. Changes in diabetes-related complications in the United States, 1990-2010. *The New England journal of medicine*. 2014;370(16):1514-1523.
21. Shrestha SS, Thompson TJ, Kirtland KA, et al. Changes in Disparity in County-Level Diagnosed Diabetes Prevalence and Incidence in the United States, between 2004 and 2012. *PLoS One*. 2016;11(8):e0159876.
22. Beckles GL, Chou C-F. Disparities in the Prevalence of Diagnosed Diabetes — United States, 1999–2002 and 2011–2014. *Morbidity and Mortality Weekly Report*. 2016;65(45):1265-1269.
23. Bowlin SJ, Morrill BD, Nafziger AN, Jenkins PL, Lewis C, Pearson TA. Validity of cardiovascular disease risk factors assessed by telephone survey: The behavioral risk factor survey. *J Clin Epidemiol*. 1993;46(6):561-571.
24. Merrill RM, Richardson JS. Validity of self-reported height, weight, and body mass index: findings from the National Health and Nutrition Examination Survey, 2001-2006. *Prev Chronic Dis*. 2009;6(4):A121-A121.
25. Office of the National Coordinator for Health Information Technology. National Trends in Hospital and Physician Adoption of Electronic Health Records. <https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records>.
26. Perlman SE. Use and Visualization of Electronic Health Record Data to Advance Public Health. *Am J Public Health*. 2021;111(2):180-182.
27. Scott KA, Bacon E, Kraus EM, et al. Evaluating Population Coverage in a Regional Distributed Data Network: Implications for Electronic Health Record–Based Public Health Surveillance. *Public Health Rep*. 2020;135(5):621-630.
28. 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;42(11):214.
29. Hirsch AG, Conderino S, Crume T, et al. Utilizing Electronic Health Records to Enhance Surveillance of Diabetes in Children, Adolescents, and Young Adults: A Study Protocol for the DiCAYA Network [Manuscript submitted for publication]. *Department of Population Health Sciences, Geisinger*. 2023.

30. Kraus EM, Brand B, Hohman KH, Baker EL. New Directions in Public Health Surveillance: Using Electronic Health Records to Monitor Chronic Disease. *J Public Health Manag Pract.* 2022;28(2).
31. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc.* 2012;19(2):212-218.
32. Upadhyaya SG, Murphree Jr DH, Ngufor CG, et al. Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes.* 2017;1(1):100-110.
33. Wilke RA, Berg RL, Peissig P, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin Med Res.* 2007;5(1):1-7.
34. Brooks C, Stephens J, Price D, et al. Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. *Primary care diabetes.* 2009;3(4):245-248.
35. 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(1):1266.
36. Klompas M, Cocoros NM, Menchaca JT, et al. State and Local Chronic Disease Surveillance Using Electronic Health Record Systems. *Am J Public Health.* 2017;107(9):1406-1412.
37. Zhong VW, Obeid JS, Craig JB, et al. An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. *J Am Med Inform Assoc.* 2016;23(6):1060-1067.
38. Tasker RC. Why Everyone Should Care About "Computable Phenotypes". *Pediatric critical care medicine : a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies.* 2017;18(5):489-490.
39. Raebel MA, Schroeder EB, Goodrich G, et al. *Validating type 1 and type 2 diabetes mellitus in the mini-sentinel distributed database using the surveillance, prevention, and management of diabetes mellitus (supreme-dm) datalink.* 2016.
40. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annual Review of Biomedical Data Science.* 2018;1(1):53-68.
41. Diagnosis and Classification of Diabetes Mellitus. *Diabetes care.* 2011;34(Supplement 1):S62-S69.
42. Ho ML, Lawrence N, van Walraven C, et al. The accuracy of using integrated electronic health care data to identify patients with undiagnosed diabetes mellitus. *J Eval Clin Pract.* 2012;18(3):606-611.
43. Committee ADAPP. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022. *Diabetes care.* 2021;45(Supplement_1):S17-S38.
44. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care.* 2013;36(4):914-921.
45. Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA. Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data. *Pharmacoepidemiol Drug Saf.* 2018;27(10):1053-1059.

46. Chi GC, Li X, Tartof SY, Slezak JM, Koebnick C, Lawrence JM. Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. *BMJ open diabetes research & care*. 2019;7(1):e000547.
47. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc*. 2016;23(6):1046-1052.
48. Pacheco JA, Rasmussen LV, Kiefer RC, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. *J Am Med Inform Assoc*. 2018;25(11):1540-1546.
49. Partners Healthcare. Type 2 Diabetes - PRS Evaluation. 2021. <https://phekb.org/phenotype/1546>.
50. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc*. 2016;23(e1):e20-27.
51. Wells BJ, Lenoir KM, Wagenknecht LE, et al. Detection of Diabetes Status and Type in Youth Using Electronic Health Records: The SEARCH for Diabetes in Youth Study. *Diabetes care*. 2020;43(10):2418-2425.
52. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina*. 2021;57(5):503.
53. McVeigh KH, Lurie-Moroni E, Chan PY, et al. Generalizability of Indicators from the New York City Macroscopic Electronic Health Record Surveillance System to Systems Based on Other EHR Platforms. *EGEMS (Washington, DC)*. 2017;5(1):25-25.
54. Spratt SE, Pereira K, Granger BB, et al. Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *Journal of the American Medical Informatics Association : JAMIA*. 2017;24(e1):e121-e128.
55. 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. *Current Epidemiology Reports*. 2017;4(4):346-352.
56. Panozzo CA, Woodworth TS, Welch EC, et al. Early impact of the ICD-10-CM transition on selected health outcomes in 13 electronic health care databases in the United States. *Pharmacoepidemiol Drug Saf*. 2018;27(8):839-847.
57. Zhong VW, Pfaff ER, Beavers DP, et al. Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study. *Pediatr Diabetes*. 2014;15(8):573-584.
58. Teltsch DY, Fazeli Farsani S, Swain RS, et al. Development and validation of algorithms to identify newly diagnosed type 1 and type 2 diabetes in pediatric population using electronic medical records and claims data. *Pharmacoepidemiol Drug Saf*. 2019;28(2):234-243.
59. Atella V, Piano Mortari A, Kopinska J, et al. Trends in age-related disease burden and healthcare utilization. *Aging cell*. 2019;18(1):e12861.
60. Lau JS, Adams SH, Boscardin WJ, Irwin CE, Jr. Young adults' health care utilization and expenditures prior to the Affordable Care Act. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*. 2014;54(6):663-671.

61. Olson DE, Rhee MK, Herrick K, Ziemer DC, Twombly JG, Phillips LS. Screening for diabetes and pre-diabetes with proposed A1C-based diagnostic criteria. *Diabetes Care*. 2010;33(10):2184-2189.
62. Ford CN, Leet RW, Kipling LM, et al. Racial differences in performance of HbA1c for the classification of diabetes and prediabetes among US adults of non-Hispanic black and white race. *Diabetic Medicine*. 2019;36(10):1234-1242.
63. Lawrence JM, Black MH, Zhang JL, et al. Validation of pediatric diabetes case identification approaches for diagnosed cases by using information in the electronic health records of a large integrated managed health care organization. *Am J Epidemiol*. 2014;179(1):27-38.
64. Wiese AD, Roumie CL, Buse JB, et al. Performance of a computable phenotype for identification of patients with diabetes within PCORnet: The Patient-Centered Clinical Research Network. *Pharmacoepidemiol Drug Saf*. 2019;28(5):632-639.
65. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. In. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020.
66. Nair S, Hsu D, Celi LA. Challenges and Opportunities in Secondary Analyses of Electronic Health Record Data. In: *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing; 2016:17-26.
67. Perlman SE. Use and visualization of electronic health record data to advance public health. In. Vol 111: American Public Health Association; 2021:180-182.
68. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. *Journal of medical systems*. 2018;42(11):1-16.
69. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
70. Little RJ, Rubin DB. *Statistical analysis with missing data*. Vol 793: John Wiley & Sons; 2019.
71. Nandram B, Choi JW. Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data. *Survey Methodology*. 2005;31(1):73-84.
72. Chen T, Li W, Zambarano B, Klompas M. Small-area estimation for public health surveillance using electronic health record data: reducing the impact of underrepresentation. *BMC Public Health*. 2022;22(1):1515.
73. Flood TL, Zhao YQ, Tomayko EJ, Tandias A, Carrel AL, Hanrahan LP. Electronic health records and community health surveillance of childhood obesity. *American journal of preventive medicine*. 2015;48(2):234-240.
74. Gelman A, Lax J, Phillips J, Gabry J, Trangucci R. Using multilevel regression and poststratification to estimate dynamic public opinion. *Unpublished manuscript, Columbia University*. 2016;2.
75. Mercer A, Lau A, Kennedy C. For weighting online opt-in samples, what matters most? 2018.
76. Wang W, Rothschild D, Goel S, Gelman A. Forecasting elections with non-representative polls. *International Journal of Forecasting*. 2015;31(3):980-991.
77. Matei A. On some reweighting schemes for nonignorable unit nonresponse. *The Survey Statistician*. 2018;77:21-33.

78. Imai K, Khanna K. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*. 2016;24(2):263-272.
79. Avramovic S, Alemi F, Kanchi R, et al. US veterans administration diabetes risk (VADR) national cohort: cohort profile. *BMJ Open*. 2020;10(12):e039489. <https://doi.org/10.1136/bmjopen-2020-039489>. Accessed 2020/12//.
80. Ruggles S, Flood S, Goeken R, Schouweiler M, Sobek M. IPUMS USA. In: IPUMS, ed. 12.0 ed. Minneapolis, MN2022.
81. Lumley T. Analysis of complex survey samples. *Journal of statistical software*. 2004;9:1-19.
82. Gelman A, Little TC. Poststratification into many categories using hierarchical logistic regression. 1997.
83. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:14065823*. 2014.
84. New York City Department of Health and Mental Hygiene. Community Health Survey Restricted Dataset. In:2015-2020.
85. Battaglia MP, Hoaglin DC, Frankel MR. Practical considerations in raking survey data. *Survey Practice*. 2009;2(5):2953.
86. Butticé MK, Highton B. How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? *Political Analysis*. 2017;21(4):449-467.
87. Downes M, Gurrin LC, English DR, et al. Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. *American journal of epidemiology*. 2018;187(8):1780-1790.
88. Casey JA, Pollak J, Glymour MM, Mayeda ER, Hirsch AG, Schwartz BS. Measures of SES for Electronic Health Record-based Research. *American journal of preventive medicine*. 2018;54(3):430-439.
89. Bhavsar NA, Gao A, Phelan M, Pagidipati NJ, Goldstein BA. Value of Neighborhood Socioeconomic Status in Predicting Risk of Outcomes in Studies That Use Electronic Health Record Data. *JAMA network open*. 2018;1(5):e182716.
90. 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. *American journal of preventive medicine*. 2019;57(6, Supplement 1):S65-S73.
91. Chunara R, Zhao Y, Chen J, et al. Telemedicine and healthcare disparities: a cohort study in a large healthcare system in New York City during COVID-19. *Journal of the American Medical Informatics Association : JAMIA*. 2021;28(1):33-41.
92. Beyond the COVID Pandemic, Telemedicine, and Health Care. *Telemedicine and e-Health*. 2020;26(11):1310-1313.
93. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*. 2022;78(1):214-226.
94. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science*. 2018;1:53.
95. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina (Kaunas, Lithuania)*. 2021;57(5).

96. Grace YY, Delaigle A, Gustafson P. *Handbook of Measurement Error Models*. CRC Press; 2021.
97. Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. *Current epidemiology reports*. 2018;5(4):343-356.
98. Sayon-Orea C, Moreno-Iribas C, Delfrade J, et al. Inverse-probability weighting and multiple imputation for evaluating selection bias in the estimation of childhood obesity prevalence using data from electronic health records. *BMC medical informatics and decision making*. 2020;20(1):9.
99. Lyles RH, Tang L, Superak HM, et al. Validation Data-based Adjustments for Outcome Misclassification in Logistic Regression: An Illustration. *Epidemiology (Cambridge, Mass)*. 2011;22(4):589-597.
100. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health services research*. 2008;43(4):1424-1441.
101. Hernán MA, Cole SR. Invited Commentary: Causal Diagrams and Measurement Bias. *American Journal of Epidemiology*. 2009;170(8):959-962.
102. Bullard KM, Cowie CC, Lessem SE, et al. Prevalence of Diagnosed Diabetes in Adults by Diabetes Type - United States, 2016. *MMWR Morb Mortal Wkly Rep*. 2018;67(12):359-361.
103. Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. *Jama*. 2014;312(12):1218-1226.
104. Link CL, McKinlay JB. Disparities in the prevalence of diabetes: is it race/ethnicity or socioeconomic status? Results from the Boston Area Community Health (BACH) survey. *Ethnicity & disease*. 2009;19(3):288-292.
105. Cheng YJ, Kanaya AM, Araneta MRG, et al. Prevalence of Diabetes by Race and Ethnicity in the United States, 2011-2016. *Jama*. 2019;322(24):2389-2398.
106. Torres RM, Souza MDS, Coelho ACC, de Mello LM, Souza-Machado C. Association between Asthma and Type 2 Diabetes Mellitus: Mechanisms and Impact on Asthma Control-A Literature Review. *Canadian respiratory journal*. 2021;2021:8830439.
107. *GeoSupport for R* [computer program]. Version 16.22016.
108. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. In: U.S. Department of Health and Human Services CfDCaP, ed. Atlanta, GA2019.
109. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey Data. In: U.S. Department of Health and Human Services CfDCaP, ed. Hyattsville, MD2013-2020.
110. Antonio-Villa NE, Fernández-Chirino L, Vargas-Vázquez A, Fermín-Martínez CA, Aguilar-Salinas CA, Bello-Chavolla OY. Prevalence Trends of Diabetes Subgroups in the United States: A Data-driven Analysis Spanning Three Decades From NHANES (1988-2018). *The Journal of Clinical Endocrinology & Metabolism*. 2021;107(3):735-742.
111. Tran L, Tran P, Tran L. A cross-sectional analysis of racial disparities in US diabetes screening at the national, regional, and state level. *Journal of Diabetes and its Complications*. 2020;34(1):107478.

112. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*. 2009;9(2):69-83.
113. Hsu WC, Araneta MRG, Kanaya AM, Chiang JL, Fujimoto W. BMI Cut Points to Identify At-Risk Asian Americans for Type 2 Diabetes Screening. *Diabetes Care*. 2014;38(1):150-158.
114. Sun M, Oliwa T, Peek ME, Tung EL. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Affairs*. 2022;41(2):203-211.
115. Vanderloo SE, Johnson JA, Reimer K, et al. Validation of classification algorithms for childhood diabetes identified from administrative data. *Pediatr Diabetes*. 2012;13(3):229-234.

Appendix A

Appendix Table 1: Literature Review Results for Diabetes Computable Phenotype Definitions.

Authors	Study Population	EHR Data Source	Methods	Performance Measures	Recommended/Optimal Computable Phenotype
Chi ⁴⁶	Aged < 20 years	Integrated managed care organization (Kaiser Permanente Southern California)	Compared rule-based algorithms using different thresholds in the ratio of ICD-10-CM diagnosis codes by DM type	Sensitivity, specificity, PPV, NPV, Youden's index, accuracy via manual review	<u>T1</u> : Not T2 <u>T2</u> : Ratio of T2 to total DM codes ≥ 0.5
Ho ⁴²	Adults aged ≥ 18 years	Large, independent academic medical center in Ottawa, Canada (Ottawa Hospital)	Rule-based criteria (using ICD-10-CM diagnosis codes, medications, and lab results)	Accuracy, weighted-Kappa statistic, sensitivity, specificity PPV via manual chart review	<u>Recognized DM</u> : DM diagnosis code <i>or</i> any order for DM medication <u>Probable DM</u> : Peak serum glucose ≥ 11.1 mmol/L, no DM diagnosis or medication <u>Possible DM</u> : Peak serum glucose ≥ 7.8 & < 11.1 mmol/L, no DM diagnosis or medication
Karlson ⁴⁹	All ages	Mass General Brigham	Penalized logistic regression model (LASSO) classifier using predictors from the literature (diagnosis codes, NLP of medications, sex, race, ethnicity, birth year)	Sensitivity, specificity, PPV, NPV via manual chart review	<u>T2</u> : Prediction model including total number of encounters with any diagnosis code, count of T2 diagnosis codes, count of T1 diagnosis codes, count of DM medication prescriptions
Kho ³¹	All ages	Clinical research network (eMerge Consortium) composed of five health systems in Washington,	Rule-based criteria (using ICD-10-CM diagnosis codes, medications, and lab results)	PPV via manual review	<u>T2</u> : ≥ 1 T2 diagnosis code AND ≥ 1 T1 medication AND ≥ 1 T2 medication AND earliest T2 medication precedes earliest T1 medication; <i>or</i> ≥ 1 T2 diagnosis code AND no T1 medication AND ≥ 1 T2 medication; <i>or</i> ≥ 1 T2 diagnosis code AND no DM medication AND ≥ 1 abnormal

		Illinois, Minnesota, Tennessee, and Wisconsin			lab (Random glucose > 200 mg/dl, Fasting glucose \geq 125 mg/dl, or Hemoglobin A1c \geq 6.5%); <i>or</i> No T2 diagnosis code AND \geq 1 T2 medication AND \geq 1 abnormal lab; <i>or</i> \geq 1 T2 diagnosis code AND no T2 medication AND \geq 1 T1 medication AND \geq 2 T2 physician diagnoses
Klompas ⁴⁴	All ages	Ambulatory practice network in Massachusetts (Atrius Health)	Compared expert-defined rule-based algorithms (using ICD-9-CM diagnosis codes, medications, and lab results)	Sensitivity, specificity, PPV via manual chart review	<u>DM</u> : Hemoglobin A1c \geq 6.5% <i>or</i> fasting glucose \geq 126 mg/dL <i>or</i> prescription for insulin outside of pregnancy <i>or</i> \geq 2 DM diagnosis codes <i>or</i> \geq 1 DM medication <u>T1</u> : Ratio of T1 to T2 diagnosis codes > 0.5 AND (prescription for glucagon <i>or</i> no record of an oral or hypoglycemic other than metformin) <i>or</i> C-peptide negative <i>or</i> diabetes autoantibodies positive <i>or</i> prescription for urine acetone test strips <u>T2</u> : not T1
Lawrence ⁶³	Aged < 20 years	Integrated managed care organization (Kaiser Permanente Southern California)	Compared rule-based algorithms from the literature (using ICD-9-CM diagnosis codes, prescriptions, lab results, age)	Sensitivity, specificity, PPV, AUC, accuracy via SEARCH for Diabetes in Youth Study registry	<u>DM</u> : \geq 1 outpatient DM diagnosis code <i>or</i> prescription for insulin <u>T1</u> : \geq 1 outpatient T1 diagnosis code <u>T2</u> : Absence of an outpatient T1 diagnosis code
Raebel ³⁹	All ages	Mini-Sentinel Distributed Database	Compared rule-based algorithms from the literature (using ICD-9-CM diagnosis codes, prescriptions, lab results)	Sensitivity & PPV via SUPREME-DM diabetes registry for DM only	<u>DM</u> : > 1 inpatient diagnosis code <i>or</i> >2 of the following events (when the two events are from the same source [e.g. two outpatient diagnoses or two elevated laboratory values], they must occur on separate dates < 730 days apart): (1) outpatient diagnosis code; (2) antidiabetic medication; (3) A1c > 6.5%; (4) Fasting plasma glucose > 126 mg/dl; (5) Random plasma glucose > 200mg/dl <u>T1</u> : Ratio of T1 to T2 diagnosis codes >0.5 and (no record of an oral hypoglycemic other than

					insulin or metformin <i>or</i> prescription for glucagon) <u>T2</u> : Not T1
Schroeder ⁴⁵	Adults aged \geq 20 years	Integrated managed care organization (Kaiser Permanente Colorado)	Compared combinations of rule-based criteria from the Klompas algorithm (using ICD-9-CM and ICD-10-CM diagnosis codes, prescriptions, lab results)	Sensitivity & PPV via manual chart review	<u>T1</u> : Ratio of T1 to T2 diagnosis codes > 0.5 AND (prescription for glucagon <i>or</i> no record of an oral or hypoglycemic other than metformin) <i>or</i> C-peptide negative or diabetes autoantibodies positive
Spratt ⁵⁴	Adults aged \geq 18 years	Large, independent academic medical center in North Carolina (Duke Health System)	Compared rule-based algorithms (using ICD-9-CM diagnosis codes, medications, and lab results) from the literature	Sensitivity & specificity via manual chart review	<u>T2</u> : T2 diagnosis code <i>or</i> DM medication from an ambulatory encounter <i>or</i> ≥ 2 abnormal lab results in the past 365 days
Teltsch ⁵⁸	Aged 10 to 18 years	United States Department of Defense health system	Compared rule-based algorithms from the literature or clinical insight using variables from the EHR (lab results, BMI, and blood pressure) or medical claims (including diagnosis codes, medication dispensings)	Sensitivity, specificity, PPV, NPV via manual chart review	<u>T1</u> : No (oral glucose lowering drug (other than metformin or metformin but not insulin) <i>or</i> ≤ 2 DM diagnosis codes) <i>or</i> (insulin <i>and</i> ratio of T2 to T1 diagnosis codes < 0.8) <i>or</i> insulin pump <i>or</i> continuous glucose monitor <u>T2</u> : Oral glucose-lowering drug other than metformin or (metformin but not insulin) <i>or</i> (no (≤ 2 T2 diagnosis codes <i>or</i> (insulin <i>and</i> ratio of T2 to T1 diagnosis codes < 0.8) <i>or</i> insulin pump <i>or</i> continuous glucose monitor) <i>and</i> (ratio of T2 to T1 diagnosis codes ≥ 0.8 <i>or</i> (long-acting insulin and no short/rapid-acting insulin))
Thorpe ^{35,53}	Adults aged \geq 20 years	Ambulatory practice EHR-data network in New York (NYC MacroScope)	Compared rule-based algorithm using ICD-9-CM diagnosis codes alone to algorithm augmented with medications and lab results	Sensitivity & specificity via manual chart review	<u>T2</u> : Last hemoglobin A1c value $\geq 6.5\%$ in past two years; <i>or</i> ever diagnosed with T2; <i>or</i> prescribed T2 medication in past year

Upadhyaya ³²	All ages	Large, independent academic medical center in Minnesota (Mayo Clinic)	Compared proposed rule-based algorithm (using ICD-9-CM diagnosis codes, medications, and natural language processing of clinical notes) to rule-based algorithms from the literature	Sensitivity, specificity, PPV, NPV via manual chart review	<u>DM</u> : ≥ 1 DM diagnosis code; <i>or</i> ≥ 1 DM medication during outpatient medication reconciliation; <i>or</i> metformin AND abnormal lab result; <i>or</i> clinical note positive for DM
Vanderloo ¹¹⁵	Aged < 20 years	Single-payer health-system in Canada (National Diabetes Surveillance System)	Compared four expert-defined rule-based algorithms (using demographics and medications)	Sensitivity, specificity, PPV via diabetes registry	<u>T1</u> : age <10 years on their index date <i>or</i> Insulin and/or glucose monitoring strips (730 days from index date) <u>T2</u> : not T1
Wei ⁵⁰	All ages	Large, independent academic medical center in Tennessee (Vanderbilt University Medical Center)	Compared combinations of rule-based criteria (using ICD-9-CM diagnosis codes, medications, and natural language processing of primary clinical notes)	Sensitivity, PPV, F-score via manual chart review	<u>T1</u> : Evidence of T1 in ≥ 2 of the following components of the EHR: (1) diagnosis codes; (2) primary clinical notes; (3) medications <u>T2</u> : Evidence of T2 in ≥ 2 of the following components of the EHR: (1) diagnosis codes; (2) primary clinical notes; (3) medications
Wells ⁵¹	Aged < 20 years	Three independent children's hospitals in Ohio, Washington, and Colorado (Cincinnati Children's Hospital, Seattle Children's, and Children's Hospital Colorado)	Compared rule-based (using ICD-10-CM diagnosis codes, medications, and lab results) and multinomial logistic regression (using demographics, diagnosis codes, medications, and lab results)	Sensitivity, specificity, PPV, NPV via manual chart review	<u>DM</u> : HbA1c $\geq 6.5\%$ <i>or</i> fasting plasma glucose ≥ 126 mg/dL <i>or</i> random plasma glucose ≥ 200 mg/dL <i>or</i> ≥ 1 DM diagnosis code <i>or</i> prescription/administration of a diabetes-related medication <u>T1</u> : ≥ 2 DM diagnosis codes with the most frequently occurring diabetes type of T1 or a tie between T1 and T2 or other diabetes <u>T2</u> : ≥ 2 DM diagnosis codes with the most frequently occurring diabetes type of T2 or a tie between T2 and other diabetes

Wiese ⁶⁴	Adults aged \geq 20 years	PCORnet clinical research network (VUHS, UNC, OneFlorida, Tulane)	Rule-based criteria (using ICD-9-CM and ICD-10-CM diagnosis codes, medications, and lab results)	PPV via manual review	<u>T2</u> : Inpatient or outpatient T2 diagnosis code and DM medication within 90 days following the diagnosis date; <i>or</i> T2 diagnosis code and outpatient HbA1c \geq 6.5% within 90 days before or after diagnosis date; <i>or</i> DM medication within 90 days before or after an outpatient HbA1c \geq 6.5%
Zhong ³⁷	Aged < 20 years	Two large, independent academic medical centers in North and South Carolina (Medical University of South Carolina & University of North Carolina Health Care System)	Compared expert-defined rule-based (using ICD-9-CM diagnosis codes, medications, and lab results) and classification and regression tree (CART) algorithms (using demographics, diagnosis codes, medications, and lab results)	Sensitivity, specificity, PPV via manual chart review	<u>DM</u> : \geq 1 billing DM diagnosis codes <u>T1</u> : Ratio of T1 diagnosis codes to the sum of T1 and T2 diagnosis codes \geq 0.5 or 0.6 <u>T2</u> : No computable phenotype met surveillance criteria, recommended approach incorporating manual review
Zhong ⁵⁷	Aged < 20 years	University of North Carolina Health Care System	Compared expert-defined rule-based (using ICD-9-CM diagnosis codes, medications, and lab results)	Sensitivity, specificity, PPV via manual chart review	<u>DM</u> : Meet \geq 2 criteria: \geq 1 Hemoglobin A1c \geq 6.0%; \geq 1 fasting glucose \geq 126 mg/dL; \geq 2 random glucose \geq 200 mg/dL; \geq 1 problem list DM diagnosis codes; \geq 1 billing data DM diagnosis code; \geq 1 DM-related outpatient medication <u>T1</u> : Ratio of T1 diagnosis codes to the sum of T1 and T2 diagnosis codes \geq 0.5 <u>T2</u> : Ratio of T2 diagnosis codes to the sum of T1 and T2 diagnosis codes \geq 0.4

Abbreviations: PPV: positive predictive value; NPV: negative predictive value; DM: diabetes; T1: type 1 diabetes; T2: type 2 diabetes; HbA1c: hemoglobin A1C.

Appendix B

Sensitivity Analyses for Model Specifications:

- Main multilevel logistic regression model specification (MLRP): including fixed effects for sex, age category, and Medicaid insurance status and random effects for race/ethnicity and PUMA.
- Alternate specification 1 (MLRP – ACS): including fixed effects for sex, age category, and Medicaid insurance, random effects for race/ethnicity, race/ethnicity*sex interaction, and PUMA, and neighborhood-level fixed effects for ACS variables (% living below the federal poverty level, % with a bachelor’s degree or higher, % unemployed, % foreign-born).
- Alternate specification 2 (MLRP – CHS): including fixed effects for sex, age category, and Medicaid insurance, random effects for race/ethnicity, race/ethnicity*sex interaction, and PUMA, and neighborhood-level fixed effects for ACS variables and NYC CHS variables (adult diabetes prevalence, adult obesity prevalence, and % of adults with a primary care physician).

Sensitivity Analyses for NYU Catchment Area Definitions:

- Geographic Definition (main text analyses): a public health-relevant approach including all PUMAs within the New York City boundaries (n = 55).

- Geographic & Penetrance Definition: a hybrid public health-relevant/data-driven approach including all PUMAS within New York City Counties with >5% penetrance (excluding Bronx County) (n = 45).
- Adjacent Neighborhood Definition: a data-driven approach including all PUMAS with >10% penetrance and contiguous PUMAs (n = 37).
- Data Penetrance Definition: a data-driven approach including all PUMAS with >10% penetrance (n = 29).

Appendix Table 2: Demographic Profile of the NYU Sample and General Population under Different Catchment Area Definitions, NYC Young Adults Aged 18-44 Years.

	Geographic ^a		Geographic & Penetrance ^b		Adjacent Neighborhoods ^c		EHR Penetrance ^d	
	Pop.	Samp.	Pop.	Samp.	Pop.	Samp.	Pop.	Samp.
Sex								
Female	51.2%	62.2%	51.3%	62.1%	51.2%	62.2%	51.6%	62.1%
Male	48.8%	37.8%	48.7%	37.9%	48.8%	37.8%	48.4%	37.9%
Race								
Black	20.3%	12.7%	18.7%	12.2%	17.7%	10.8%	14.3%	8.9%
Latino	29.6%	19.1%	23.9%	18.1%	22.3%	17.9%	17.6%	16.3%
Other	18.1%	16.1%	20.5%	16.4%	20.1%	16.3%	20.0%	15.9%
White	32.0%	52.1%	36.8%	53.4%	39.8%	55.0%	48.1%	58.9%
Age								
18-29	43.6%	37.5%	42.9%	37.6%	42.4%	37.8%	41.6%	37.8%
30-44	56.4%	62.5%	57.1%	62.4%	57.6%	62.2%	58.4%	62.2%
Insurance								
Non-Medicaid	74.2%	77.8%	77.5%	77.8%	77.8%	77.5%	80.5%	78.2%
Medicaid	25.8%	22.2%	22.5%	22.2%	22.2%	22.5%	19.5%	21.8%

^a *Geographic Definition: includes all PUMAs within the New York City boundaries (n = 55).*

^b *Geographic & Penetrance Definition: includes all PUMAs within New York City Counties with >5% penetrance (excludes Bronx County) (n = 45).*

^c *Adjacent Neighborhood Definition: includes all PUMAs with >10% penetrance and contiguous PUMAs (n = 37).*

^d *Data Penetrance Definition: includes all PUMAs with >10% penetrance (n = 29).*

Abbreviations: Pop. = general population; Samp. = EHR sample.

Appendix Table 3: Overall Diabetes Prevalence Estimates (and 95% CIs) under Different Catchment Area Definitions, NYC Young Adults Aged 18-44 Years.

	Geographic^a	Geographic & Penetrance^b	Adjacent Neighborhoods^c	Data Penetrance^d
Inference to the NYC General Populationⁱ				
Gold Standard ^e	3.33% (3.02-3.67)	3.33% (3.02-3.67)	3.33% (3.02-3.67)	3.33% (3.02-3.67)
Crude	3.09% (3.04-3.14)	3.01% (2.96-3.07)	2.98% (2.93-3.04)	2.91% (2.86-2.96)
Raking	3.55% (3.46-3.63)	3.30% (3.24-3.37)	3.31% (3.24-3.37)	3.30% (3.23-3.38)
Poststratification	3.54% (3.43-3.64)	3.30% (3.24-3.36)	3.30% (3.24-3.37)	3.30% (3.23-3.38)
MLRP ^f	3.55% (3.47-3.63)	3.30% (3.24-3.36)	3.28% (3.21-3.35)	3.24% (3.17-3.32)
MLRP – ACS ^g	3.59% (3.51-3.67)	3.41% (3.34-3.49)	3.38% (3.31-3.47)	3.33% (3.26-3.41)
MLRP – CHS ^h	3.58% (3.50-3.66)	3.40% (3.33-3.48)	3.38% (3.30-3.45)	3.34% (3.25-3.42)
Inference to General Population from In-Sample Neighborhoods^j				
Gold Standard ^e	3.33% (3.02-3.67)	3.09% (2.76-3.46)	2.90% (2.56-3.29)	2.47% (2.13-2.88)
Crude	3.09% (3.04-3.14)	3.01% (2.96-3.07)	2.98% (2.93-3.04)	2.91% (2.86-2.96)
Raking	3.55% (3.46-3.63)	3.17% (3.11-3.23)	3.14% (3.07-3.20)	2.97% (2.91-3.03)
Poststratification	3.54% (3.43-3.64)	3.16% (3.09-3.23)	3.11% (3.04-3.18)	2.96% (2.89-3.03)
MLRP ^f	3.55% (3.47-3.63)	3.19% (3.13-3.25)	3.15% (3.08-3.22)	2.99% (2.92-3.04)
MLRP – ACS ^g	3.59% (3.51-3.67)	3.20% (3.13-3.26)	3.16% (3.09-3.22)	2.99% (2.93-3.05)
MLRP – CHS ^h	3.58% (3.50-3.66)	3.20% (3.14-3.25)	3.16% (3.09-3.22)	2.99% (2.92-3.04)

^a *Geographic Definition: includes all PUMAs within the New York City boundaries (n = 55).*

^b *Geographic & Penetrance Definition: includes all PUMAs within New York City Counties with >5% penetrance (excludes Bronx County) (n = 45).*

^c *Adjacent Neighborhood Definition: includes all PUMAs with >10% penetrance and contiguous PUMAs (n = 37).*

^d *Data Penetrance Definition: includes all PUMAs with >10% penetrance (n = 29).*

^e *Gold standard prevalence estimates from NYC Community Health Survey 2015-2020 data.*

^f *Multilevel logistic regression model including fixed effects for sex, age category, and Medicaid insurance status, random effects for race/ethnicity and PUMA*

^g *Multilevel logistic regression model including fixed effects for sex, age category, and Medicaid insurance, random effects for race/ethnicity, race/ethnicity*sex interaction, and PUMA, and neighborhood-level fixed effects for ACS variables (% living below the federal poverty level, % with a bachelor's degree or higher, % unemployed, % foreign-born).*

^h *Multilevel logistic regression model including fixed effects for sex, age category, and Medicaid insurance, random effects for race/ethnicity, race/ethnicity*sex interaction, and PUMA, and neighborhood-level fixed effects for ACS variables and NYC CHS variables (adult diabetes prevalence, adult obesity prevalence, and % of adults with a primary care physician).*

ⁱ *Inference to the citywide NYC general population. Raking and post-stratification “Geographic & Penetrance”, “Adjacent Neighborhoods”, and “Data Penetrance” estimates do not include PUMA in the adjustment methods due to missing strata of PUMAs in the restricted samples.*

^j *Inference to the general populations restricted to equivalent PUMAs as the sample definitions.*

Appendix Table 4: Relative Difference in EHR-Based Diabetes Prevalence Estimates from Gold Standard under Different Catchment Area Definitions, NYC Young Adults Aged 18-44 Years.

	Geographic^a	Geographic & Penetrance^b	Adjacent Neighborhoods^c	Data Penetrance^d
Inference to the NYC General Population^h				
Crude	-7.88%	-10.5%	-11.70%	-14.50%
Raking	6.02%*	-0.86%*	-0.82%*	-0.87%*
Poststratification	5.75%*	-1.05%*	-0.95%*	-0.91%*
MLRP ^e	6.16%*	-1.11%*	-1.50%*	-2.81%*
MLRP – ACS ^f	7.05%	2.25%*	1.51%*	-0.01%*
MLRP – CHS ^g	6.96%	1.94%*	1.26%*	0.11%*
Inference to General Population from In-Sample Neighborhoodsⁱ				
Crude	-7.88%	-2.58%*	2.69%*	14.9%
Raking	6.02%*	2.46%*	7.52%	16.6%
Poststratification	5.75%*	2.09%*	6.77%	16.3%
MLRP ^e	6.16%*	3.11%*	7.81%	17.1%
MLRP – ACS ^f	7.05%	3.40%*	8.02%	17.1%
MLRP – CHS ^g	6.96%	3.34%*	8.04%	17.1%

^a *Geographic Definition: includes all PUMAs within the New York City boundaries (n = 55).*

^b *Geographic & Penetrance Definition: includes all PUMAs within New York City Counties with >5% penetrance (excludes Bronx County) (n = 45).*

^c *Adjacent Neighborhood Definition: includes all PUMAs with >10% penetrance and contiguous PUMAs (n = 37).*

^d *Data Penetrance Definition: includes all PUMAs with >10% penetrance (n = 29).*

^e *Multilevel logistic regression model including fixed effects for sex, age category, and Medicaid insurance status, random effects for race/ethnicity and PUMA*

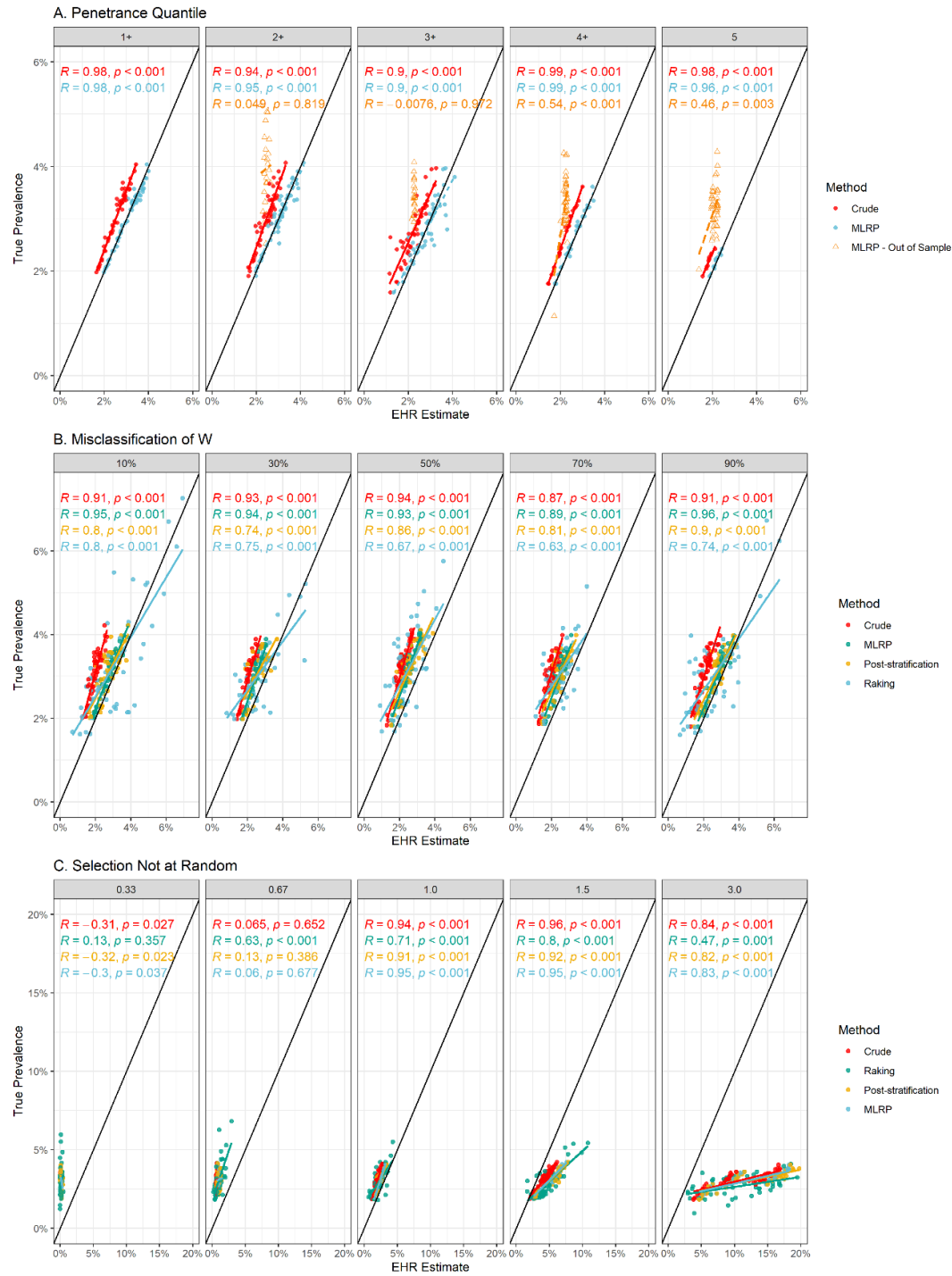
^f *Multilevel logistic regression model including fixed effects for sex, age category, and Medicaid insurance, random effects for race/ethnicity, race/ethnicity*sex interaction, and PUMA, and neighborhood-level fixed effects for ACS variables (% living below the federal poverty level, % with a bachelor's degree or higher, % unemployed, % foreign-born).*

^g *Multilevel logistic regression model including fixed effects for sex, age category, and Medicaid insurance, random effects for race/ethnicity, race/ethnicity*sex interaction, and PUMA, and neighborhood-level fixed effects for ACS variables and NYC CHS variables (adult diabetes prevalence, adult obesity prevalence, and % of adults with a primary care physician).*

^h *Inference to the citywide NYC general population. Raking and post-stratification “Geographic & Penetrance”, “Adjacent Neighborhoods”, and “Data Penetrance” estimates do not include PUMA in the adjustment methods due to missing strata of PUMAs in the restricted samples.*

ⁱ *Inference to the general populations restricted to equivalent PUMAs as the sample definitions.*

*Statistically Equivalent through the TOST test alpha = 0.05, equivalence bounds = 0.005.



Appendix Figure 1: Relative Bias in the Neighborhood-Level EHR-Based Estimates vs. the True Diabetes Prevalence by Simulation Scenario.⁶

⁶ Each point represents a neighborhood. Panel A: Scenario 1 modified definition of the catchment area based on quantiles of penetrance; Panel B: Scenario 2 modified the level of misclassification of the auxiliary variable W

Appendix Table 5: Coverage in Overall EHR-Based Estimates by Adjustment Method and Simulation Scenario.

Sample Inclusion Criteria	Crude	Raking	Post-Stratification	MLRP	MLRP – Out of Sample
Scenario 1^a					
Quantiles 1+	5%	44%	46%	46%	92%
Quantiles 2+	3%	44%	51%	48%	53%
Quantiles 3+	3%	36%	44%	41%	19%
Quantiles 4+	5%	33%	38%	37%	25%
Quantile 5	3%	32%	32%	32%	14%
Scenario 2^b					
10%	2%	68%	53%	65%	-
30%	1%	16%	3%	11%	-
50%	0%	9%	6%	1%	-
70%	0%	18%	0%	8%	-
90%	0%	61%	39%	62%	-
Scenario 3^c					
0.33	0%	0%	0%	0%	-
0.67	0%	0%	0%	0%	-
1.00	1%	7%	4%	6%	-
1.50	22%	0%	0%	0%	-
3.00	0%	0%	0%	0%	-

^a Scenario 1 modified definition of the catchment area based on quantiles of penetrance.

^b Scenario 2 modified the level of misclassification of the auxiliary variable *W* compared to the unobserved variable *U*.

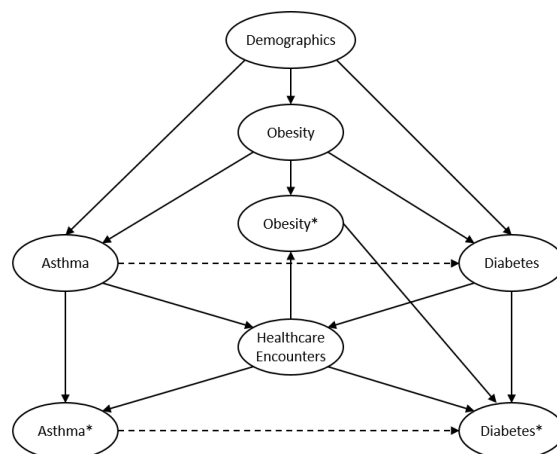
^c Scenario 3 modified the association between diabetes and selection (OR_{DM}).

compared to the unobserved variable *U*; Panel C: Scenario 3 modified the association between diabetes and selection (OR_{DM}).

Appendix C

Appendix Table 6: Case-Insensitive Search Terms for Endocrinology Review of Systems.

Endocrinology Key Terms	Review of System Key Terms
endocrin	ros
endo:	review of system
endo-	



Appendix Figure 2: Hypothesized Directed Acyclic Graph for Information Bias and Associations between Asthma and Diabetes.⁷

⁷ *Observed disease status

Appendix Table 7: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, Health Survey Estimates.

	BRFSS ^a	BRFSS – SA ^b	NHANES ^c	NHANES – SA ^d
Race/Ethnicity^e (Ref = White)				
Black	1.51 (1.3-1.75)	1.49 (1.30-1.70)	1.59 (1.18-2.14)	1.32 (0.92-1.88)
Latino	1.53 (1.32-1.77)	1.51 (1.33-1.71)	1.54 (1.14-2.07)	1.41 (0.99-2.00)
Asian	1.08 (0.81-1.45)	1.18 (0.91-1.52)	1.04 (0.71-1.52)	0.68 (0.41-1.10)
Other	1.62 (1.28-2.04)	1.64 (1.34-2.00)	1.47 (0.9-2.39)	1.16 (0.68-2.00)
Asthma^f (Ref = No)	1.23 (1.09-1.4)	1.34 (1.17-1.53)	1.38 (1.01-1.91)	1.32 (0.90-1.93)

^a Associations with self-reported diabetes diagnosis observed among Behavioral Risk Factor Surveillance System Survey 2019 respondents aged 18-44 years who report having a personal healthcare provider.

^b BRFSS sensitivity analysis: associations with self-reported diabetes diagnosis observed among Behavioral Risk Factor Surveillance System Survey 2019 respondents aged 18-44 years.

^c Associations with self-reported diabetes diagnosis or undiagnosed diabetes (A1C ≥ 6.5% or fasting glucose ≥ 126mm/Hg) observed among National Health and Nutrition Examination Survey 2013-March 2020 respondents aged 18-44 years.

^d NHANES Sensitivity analysis: with self-reported diabetes diagnosis observed among National Health and Nutrition Examination Survey 2013-March 2020 respondents aged 18-44 years.

^e ORs for race/ethnicity estimated in reference to non-Hispanic White controlling for age and sex.

^f ORs for asthma estimated in reference to non-asthmatic controlling for age, sex, race/ethnicity, obesity, insurance status, and poverty level.

Appendix Table 8: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, EHR-Based Estimates

	Naive ^a	Complete Case ^b	Missing Data ^c	Causal ^d
Race/Ethnicity^e (Ref = White)				
Black	1.79 (1.7-1.88)	1.68 (1.59-1.76)	1.73 (1.64-1.83)	1.75 (1.67-1.84)
Latino	1.93 (1.85-2.01)	1.76 (1.69-1.84)	1.75 (1.67-1.84)	1.64 (1.57-1.71)
Asian	1.11 (1.04-1.18)	1.18 (1.11-1.26)	1.26 (1.18-1.34)	1.13 (1.06-1.2)
Other	1.41 (1.3-1.51)	1.46 (1.35-1.58)	1.57 (1.45-1.7)	1.4 (1.3-1.5)
Asthma^f (Ref = No)	3.01 (2.86-3.18)	1.87 (1.77-1.97)	1.79 (1.67-1.92)	1.42 (1.34-1.51)

^a Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2018.

^b Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2018 with complete records, as defined as those with a review of systems for endocrinology.

^c Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2018 with complete records, as defined as those with a review of systems for endocrinology, using missing data framework weighting the complete case subset with stabilized IPW.

^d Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2018, using the causal framework controlling for total number of encounters.

^e ORs for race/ethnicity estimated in reference to non-Hispanic White controlling for age and sex.

^f ORs for asthma estimated in reference to non-asthmatic controlling for age, sex, race/ethnicity, obesity, insurance status, and poverty level.

Appendix Table 9: Sensitivity Analyses of Complete Case Definitions: Descriptive Summary of NYU Patient Population by Diabetes Status

	≥ 1 Endocrinology Review of Systems		≥ 1 Endocrinology & Respiratory Review of Systems		≥ 1 DM-related Lab & BMI Measurement	
	Non-Diabetic	Diabetic	Non-Diabetic	Diabetic	Non-Diabetic	Diabetic
Total	166992 (94.5)	9698 (5.5)	289203 (95.9)	12366 (4.1)	190369 (94.1)	11898 (5.9)
Age (30-44)	104005 (62.3)	7519 (77.5)	177944 (61.5)	9622 (77.8)	122123 (64.2)	9248 (77.7)
Sex (Male)	65240 (39.1)	4017 (41.4)	112878 (39.0)	5112 (41.3)	71579 (37.6)	4982 (41.9)
Medicaid	37759 (22.6)	2902 (29.9)	64665 (22.4)	3754 (30.4)	46826 (24.6)	3796 (31.9)
Raw Race/Ethnicity						
White	73996 (44.3)	3412 (35.2)	126615 (43.8)	4452 (36.0)	83192 (43.7)	4248 (35.7)
Black	18847 (11.3)	1691 (17.4)	32195 (11.1)	2053 (16.6)	21453 (11.3)	1912 (16.1)
Latino	27479 (16.5)	2501 (25.8)	46955 (16.2)	3122 (25.2)	35439 (18.6)	3098 (26.0)
Asian/PI	13227 (7.9)	672 (6.9)	22359 (7.7)	888 (7.2)	16727 (8.8)	973 (8.2)
Other	11709 (7.0)	747 (7.7)	20925 (7.2)	980 (7.9)	14891 (7.8)	972 (8.2)
Missing	21734 (13.0)	675 (7.0)	40154 (13.9)	871 (7.0)	18667 (9.8)	695 (5.8)
Imputed Race/Ethnicity						
White	86629 (51.9)	3744 (38.6)	149693 (51.8)	4872 (39.4)	94025 (49.4)	4603 (38.7)
Black	21829 (13.1)	1832 (18.9)	37656 (13.0)	2233 (18.1)	24033 (12.6)	2046 (17.2)
Latino	33696 (20.2)	2821 (29.1)	58420 (20.2)	3534 (28.6)	41883 (22.0)	3460 (29.1)
Asian/PI	16599 (9.9)	787 (8.1)	28523 (9.9)	1037 (8.4)	19889 (10.4)	1096 (9.2)
Other	8239 (4.9)	514 (5.3)	14911 (5.2)	690 (5.6)	10539 (5.5)	693 (5.8)
Any BMI	163127 (97.7)	9651 (99.5)	281030 (97.2)	12299 (99.5)	190369 (100)	11898 (100)
Obese	34593 (20.7)	4499 (46.4)	58389 (20.2)	5626 (45.5)	42429 (22.3)	5294 (44.5)
Encounters (mean (sd))*	22.27 (28.19)	55.41 (60.49)	18.59 (24.36)	49.43 (56.13)	23.09 (27.53)	50.71 (56.68)
Duration (mean (sd))	2.26 (1.97)	3.09 (2.02)	2.03 (1.95)	2.96 (2.03)	2.50 (1.98)	3.05 (2.03)
Routine Medical Exam*	77927 (46.7)	3927 (40.5)	104336 (36.1)	4456 (36.0)	98508 (51.7)	4421 (37.2)
DM-related Lab^a	112752 (67.5)	8790 (90.6)	169914 (58.8)	10933 (88.4)	190369 (100)	11898 (100)
PUMA Coverage						
< 10 %	28143 (16.9)	2215 (22.8)	49254 (17.0)	2718 (22.0)	30787 (16.2)	2543 (21.4)
10-<20%	57185 (34.2)	3244 (33.5)	96460 (33.4)	4074 (32.9)	62652 (32.9)	3813 (32.0)
20-<30%	57597 (34.5)	2808 (29.0)	101763 (35.2)	3612 (29.2)	66311 (34.8)	3564 (30.0)
30-<40%	24067 (14.4)	1431 (14.8)	41726 (14.4)	1962 (15.9)	30619 (16.1)	1978 (16.6)
Asthma (Yes)	11172 (6.7)	1563 (16.1)	15887 (5.5)	1819 (14.7)	14039 (7.4)	1854 (15.6)

^a Including all A1c, random blood glucose, and fasting blood glucose lab results.

*Defined within the years of 2017-2019.

Appendix Table 10: Sensitivity Analyses of Complete Case Definitions: Odds Ratios for Diabetes by Race/Ethnicity and Asthma, EHR-Based Estimates Complete Case and Missing Data Estimates

	≥ 1 Endocrinology Review of Systems		≥ 1 Endocrinology & Respiratory Review of Systems		≥ 1 DM-related Lab & BMI Measurement	
	Complete Case ^a	Missing Data ^b	Complete Case ^a	Missing Data ^b	Complete Case ^a	Missing Data ^b
Race/Ethnicity^c (Ref = White)						
Black	1.89 (1.78-2.01)	1.72 (1.64-1.81)	1.88 (1.78-2)	2.03 (1.89-2.18)	1.72 (1.63-1.82)	1.76 (1.66-1.87)
Latino	1.97 (1.87-2.08)	1.74 (1.67-1.82)	1.95 (1.86-2.06)	2.07 (1.94-2.2)	1.73 (1.65-1.81)	1.74 (1.65-1.83)
Asian	1.11 (1.02-1.2)	0.97 (0.91-1.04)	1.09 (1.01-1.19)	1.13 (1.03-1.24)	1.14 (1.06-1.22)	1.3 (1.21-1.39)
Other	1.38 (1.25-1.52)	1.24 (1.14-1.35)	1.36 (1.24-1.5)	1.45 (1.3-1.62)	1.31 (1.2-1.42)	1.42 (1.3-1.55)
Asthma^d (Ref = No)						
	1.83 (1.71-1.96)	1.73 (1.66-1.81)	2.23 (2.1-2.37)	2.28 (2.1-2.46)	2.06 (1.95-2.17)	1.98 (1.85-2.12)

^a Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2018 with complete records, as defined by column header.

^b Associations with EHR-defined diabetes status observed among NYC resident NYU patient population with an inpatient or outpatient encounter from 2017-2018 with complete records, as defined by column header.

^c ORs for race/ethnicity estimated in reference to non-Hispanic White controlling for age and sex.

^d ORs for asthma estimated in reference to non-asthmatic controlling for age, sex, race/ethnicity, obesity, Medicaid insurance status, and poverty level.