Appendix I

Data Illustration

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 Service 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 1: EHR-Based Diabetes Prevalence Estimates by Demographic Subgroups, NYC Young Adults Aged 18-44 Years.

	Crude	Raking	Post-Stratification	MLRP
Sex				
Female	2.93%	3.41%	3.36%	3.34%
Male	3.35%	3.69%	3.72%	3.83%
Race				
Black	4.23%	4.50%	4.43%	4.38%
White	2.38%	2.45%	2.44%	2.43%
Age				
18-29	1.88%	2.13%	2.25%	2.29%
30-44	3.82%	4.64%	4.53%	4.58%

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

	Geographic[a]	Geographic & Penetrance[b]	Adjacent Neighborhoods[c]	Data Penetrance[d]
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).*

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

	Geographic[a]	Geographic & Penetrance[b]	Adjacent Neighborhoods[c]	Data Penetrance[d]
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).*
**Reject the null hypothesis of the TOST, or equivalent to the gold standard within equivalence bounds of 0.005.*

Simulations

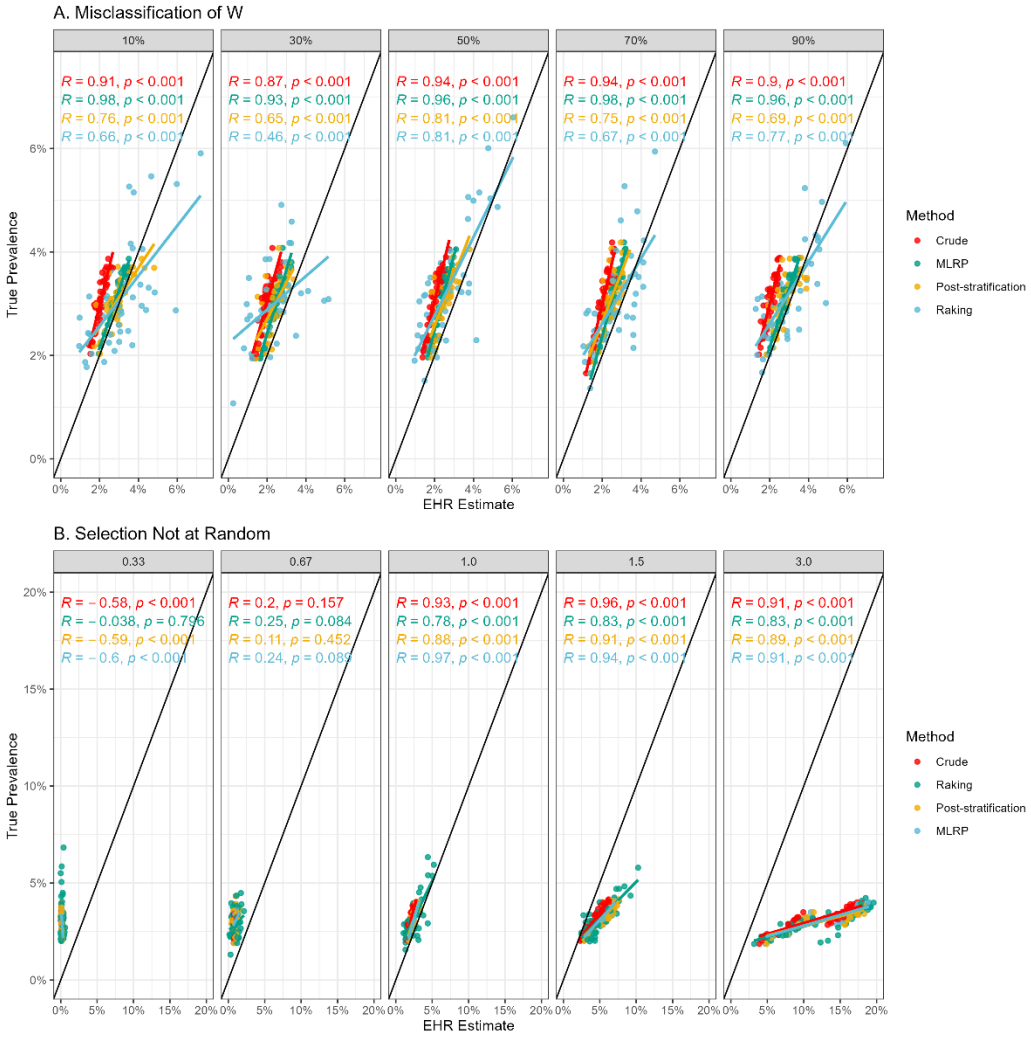
Simulation Scenario 1:

- Selection model (individual *i* in Sex∗Race group *k*):
  - $\text{Logit}(\text{odds}_{\text{selection}}) = -0.11 - 0.60\text{Race}_{\text{NHB}} - 0.34\text{Race}_{\text{HIS}} - 0.54\text{Race}_{\text{OTH}} - 0.13\text{Age}_{18\text{-}29} - 0.05\text{Distance}_1 - 0.60\text{Distance}_2 - 1.39\text{Distance}_3 - 0.39\text{Sex}_{\text{male}} - 0.36U_1 + u_k$
  - $u_k \sim N(0,0.3)$
- Diabetes model (individual *i* in neighborhood cluster *j*):
  - $\text{Logit}(\text{odds}_{\text{DM}}) = -3.91 + 1.32\text{Age}_{30\text{-}44} - 0.49\text{Sex}_{\text{female}} + 0.59\text{Race}_{\text{NHB}} + 0.81\text{Race}_{\text{HIS}} + 0.52\text{Race}_{\text{OTH}} + 0.41\text{Sex}*\text{Race}_{\text{female,NHB}} + 0.69U + u_j$
  - $u_j \sim N(0,0.5)$
- *U* model (individual *i*):
  - $\text{Logit}(\text{odds}_U) = -0.36 + 0.05\text{Sex}_{\text{female}} + 0.13\text{Race}_{\text{NHB}} + 0.21\text{Race}_{\text{HIS}} + 0.15\text{Race}_{\text{OTH}}$

Simulation Scenario 2:

- Selection model (individual *i* in Sex∗Race group *k*):
  - $\text{Logit}(\text{odds}_{\text{selection}}) = -0.11 - 0.60\text{Race}_{\text{NHB}} - 0.34\text{Race}_{\text{HIS}} - 0.54\text{Race}_{\text{OTH}} - 0.13\text{Age}_{18\text{-}29} - 0.05\text{Distance}_1 - 0.60\text{Distance}_2 - 1.39\text{Distance}_3 - 0.39\text{Sex}_{\text{male}} - \beta_1 DM_1 + u_k$
  - $u_k \sim N(0,0.3)$
    - β_1 modified at the levels of 0.33, 0.67, 1.0, 1.5, and 3.0
- Diabetes model (individual *i* in neighborhood cluster *j*):
  - $\text{Logit}(\text{odds}_{\text{DM}}) = -3.91 + 1.32\text{Age}_{30\text{-}44} - 0.49\text{Sex}_{\text{female}} + 0.59\text{Race}_{\text{NHB}} + 0.81\text{Race}_{\text{HIS}} + 0.52\text{Race}_{\text{OTH}} + 0.41\text{Sex}*\text{Race}_{\text{female,NHB}} + 0.69U + u_j$
  - $u_j \sim N(0,0.5)$
- *U* model (individual *i*):
  - $\text{Logit}(\text{odds}_U) = -0.36 + 0.05\text{Sex}_{\text{female}} + 0.13\text{Race}_{\text{NHB}} + 0.21\text{Race}_{\text{HIS}} + 0.15\text{Race}_{\text{OTH}}$

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



Each point represents a neighborhood. Panel A: Scenario 1 modified the level of misclassification of the auxiliary variable *W* compared to the unobserved variable *U*; Panel B: Scenario 2 modified the association between diabetes and selection (OR_{DM}).

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

Sample Inclusion Criteria	Crude	Raking	Post-Stratification	MLRP
Scenario 1[a]				
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 2[b]				
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%
0.33	0%	0%	0%	0%

[a] *Scenario 1 modified the level of misclassification of the auxiliary variable W compared to the unobserved variable U.*

[b] *Scenario 2 modified the association between diabetes and selection (OR_{DM}).*