Adjusting for Measurement Error to Quantify the Relationship Between Diabetes and Local Access to Healthy Food

Wake Forest University
Department of Statistical Sciences

Ashley E. Mullan

March 2024

Follow along with me!

https://bit.ly/ashley_talks

SCAN ME

Motivation

Healthy Eating ➡️ Healthy Living

- A healthy diet increases the likelihood of good overall health and **decreases risk of preventable illness** (World Health Organization, 2019).
- Maintaining a healthy diet requires **consistent access to healthy food**, which may be hindered by geography or income.
- Review studies found **high prevalence of diabetes** in food-insecure households (Gucciardi et al., 2014).

Measuring Food Access

- Count the number of healthy food retailers in a given radius (i.e., **density**)
- Compute the distance to the nearest healthy food retailer (i.e., **proximity**)
- Create an **indicator** of "low" food access that evaluates to 1 if zero healthy food retailers exist within a given distance (e.g., 0.5 miles or 1 mile).

Distance Computations

- The **Haversine distance** is a trigonometric function of latitude and longitude.
- It ignores physical obstacles, so it **underestimates** the true distance between two points and is considered **error-prone**.
- The Haversine distance in the image is **impassable**, as it crosses a pond.

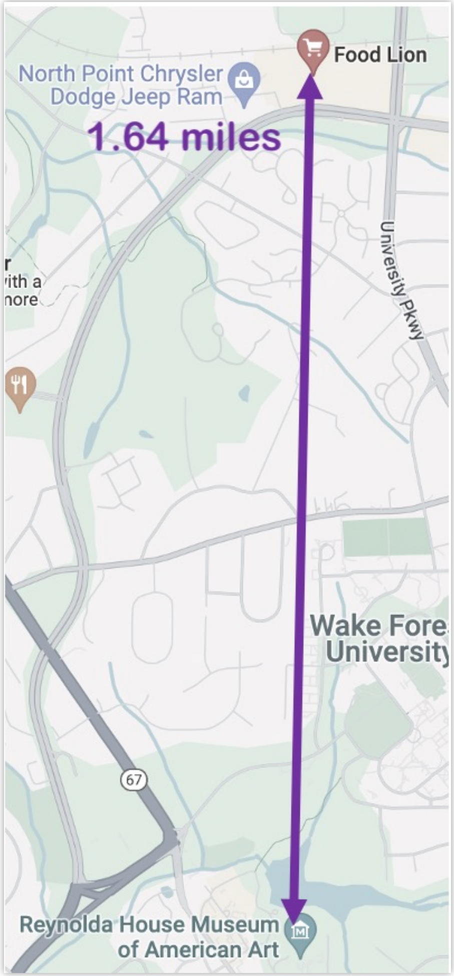

Figure: Haversine distance from Reynolda Manor House to a nearby Food Lion

Distance Computations

- The **route-based distance** works around obstacles.
- It is **more accurate** than the Haversine distance but is **computationally expensive**.

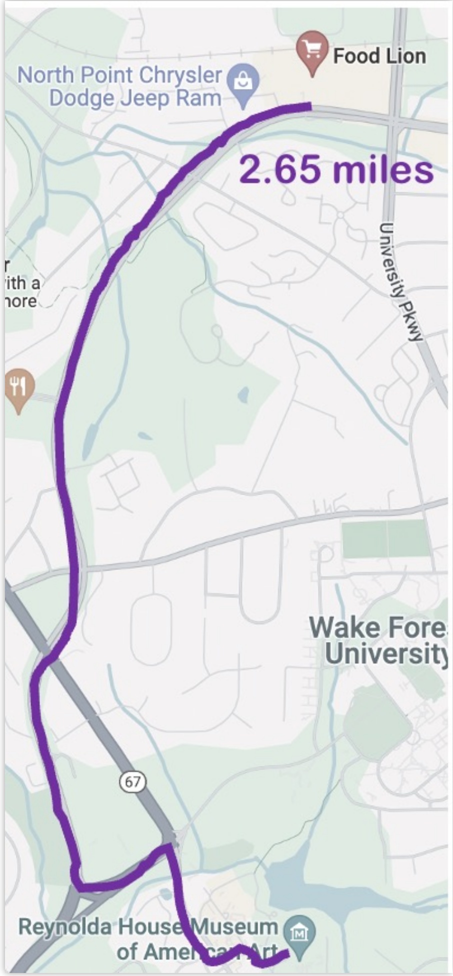

Figure: Route distance from Reynolda Manor House to a nearby Food Lion

Research Questions

1. Can we use a function of distance to healthy food retailers to **quantify food access** in the Piedmont area of North Carolina, even if this function is **subject to measurement error**?
2. Can we estimate the relationship between **low food access** and **diabetes** prevalence?

Methods

Notation

- X is an error-free binary explanatory variable for low food access based on route-based distances
- X^* is an error-prone version of X based on Haversine distances
- Z is an error-free covariate vector
- Y is a count of diabetes cases in the area of interest
- Q is an indicator of whether an observation has been queried

We want to estimate the coefficients β from the Poisson model of $\mathrm{Y} \mid \mathrm{X}, \mathrm{Z}$.

Two-Phase Design

- Having **some correct** route-based distances is better than none.
- Error-prone Haversine distances are available for all N neighborhoods, and we can use them to create our indicator of low food access X^* that is subject to misclassification.
- In addition to X^*, we **query** route-based distances to create our indicator X for n neighborhoods, where $n < N$.

Figure: An example of two-phase design.

Modeling Options

- **Gold Standard**
- Naïve Regression
- Complete Case Analysis
- Maximum Likelihood Estimation

👍

This method achieves optimal bias and variance.

👎

This method assumes we have all of the correct data available.

Modeling Options

- Gold Standard
- **Naïve Regression**
- Complete Case Analysis
- Maximum Likelihood Estimation

👍

The model is easy to fit and utilizes information from the error-prone data for all N neighborhoods.

👎

The model is biased by a function of the sensitivity and specificity (Shaw et al., 2020).

Modeling Options

- Gold Standard
- Naïve Regression
- **Complete Case Analysis**
- Maximum Likelihood Estimation

👍

The model is unbiased, as it uses the error-free measurements.

👎

The model does not take the unqueried data into account.

Modeling Options

- Gold Standard
- Naïve Regression
- Complete Case Analysis
- **Maximum Likelihood Estimation**

👍

The model utilizes information from both the queried and unqueried observations.

👎

This method is not (yet) implemented in existing software.

More on the MLE

$$\ell(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{i=1}^{N} Q_i \log P_{\boldsymbol{\beta},\boldsymbol{\eta}}(X, X^*, Y, \boldsymbol{Z}) + (1 - Q_i) \log P_{\boldsymbol{\beta},\boldsymbol{\eta}}(Y, X^*, \boldsymbol{Z})$$

More on the MLE

Poisson error

$$P(Y, X, \boldsymbol{Z}, X^*) = \overbrace{P(Y \mid X, X^*, \boldsymbol{Z})}\overbrace{P(X \mid X^*, \boldsymbol{Z})}P(X^*, \boldsymbol{Z})$$

$$= P_{\boldsymbol{\beta}}(Y \mid X, \boldsymbol{Z})P(X \mid X^*, \boldsymbol{Z})P(X^*, \boldsymbol{Z})$$

$$\propto P_{\boldsymbol{\beta}}(Y \mid X, \boldsymbol{Z})P(X \mid X^*, \boldsymbol{Z})$$

$$P(Y, X^*, \boldsymbol{Z}) = \sum_{x=0}^{1} P(Y, X = x, \boldsymbol{Z}, X^*)$$

Simulations

Roadmap

We **vary**:

- Sample size N
- Queried sample size n
- Error mechanism

We **compare**:

- Gold Standard
- Complete Case
- Naïve Model
- MLE

We **observe** the effect of interest $\widehat{\beta_1}$ (truth = 0.155) and the relative efficiency.

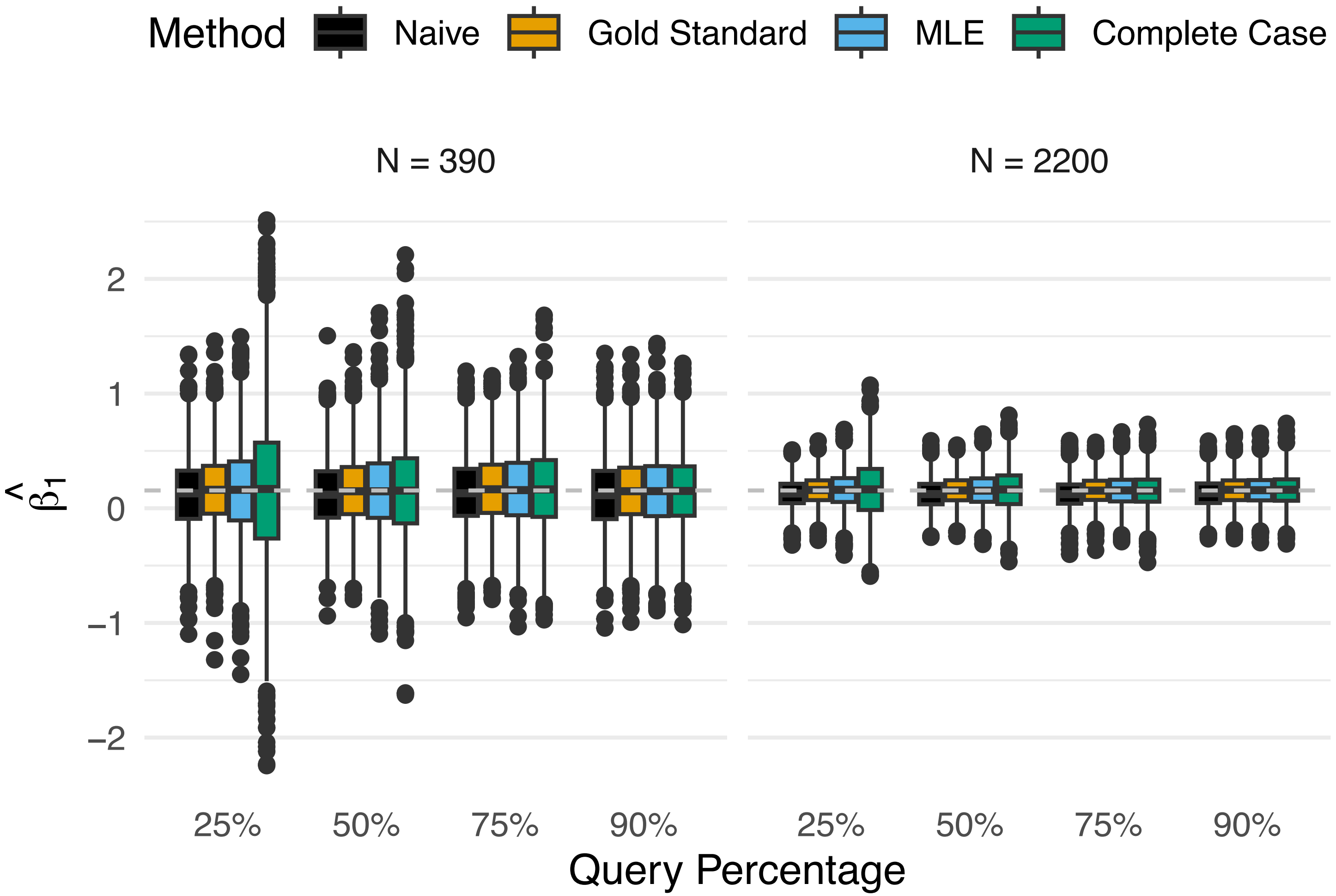

Figure: Box plot comparing method performance across different query percentages

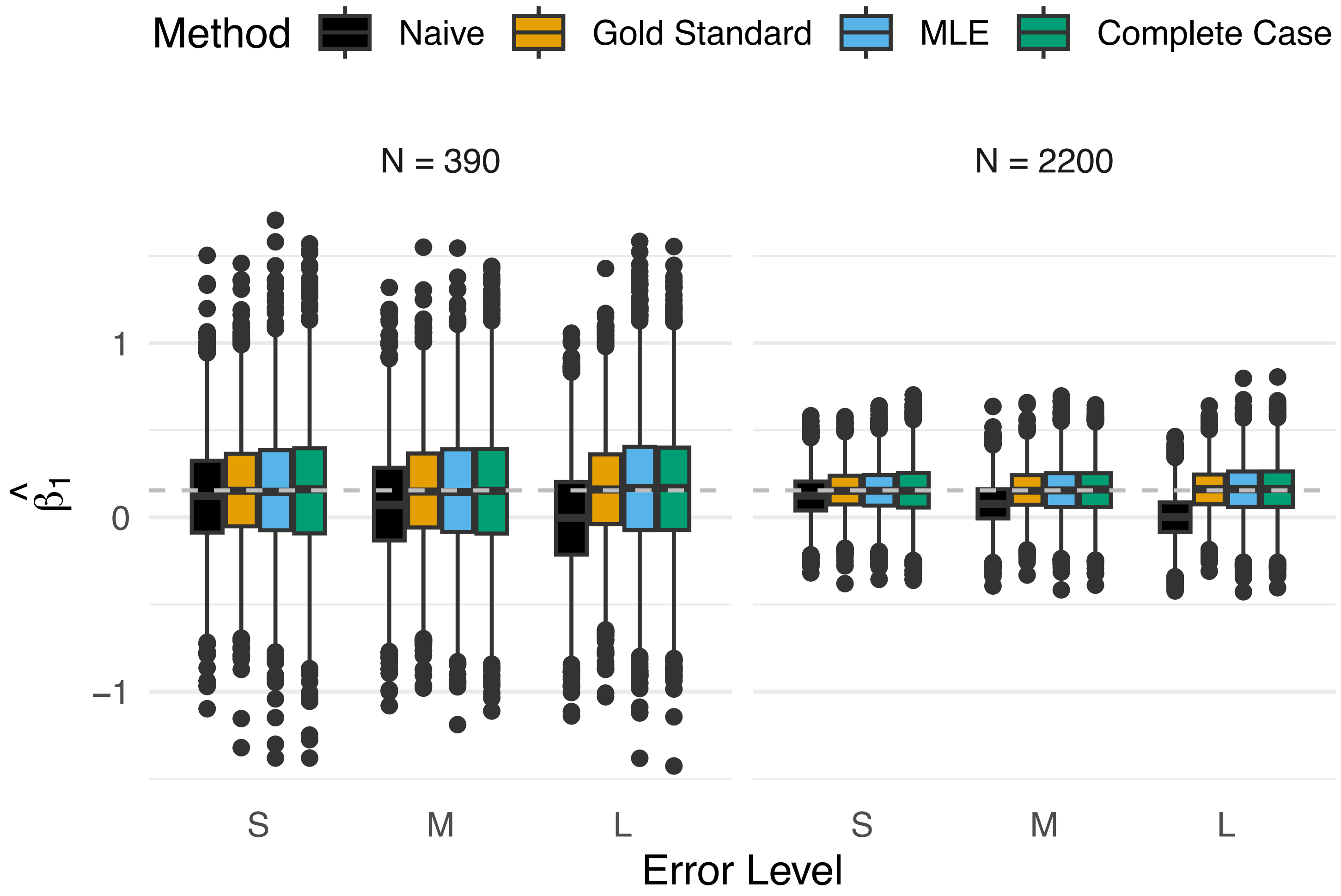

Figure: Box plot comparing method performance across different error settings

Summary

- Across all four query settings, the MLE remains **fairly unbiased**.
- As we vary the size of the queried sample n, the MLE recovers up to 91% of the **efficiency** of the gold standard model and beats the complete case model in every case.
- As we introduce more error into the input data, the MLE remains **fairly unbiased**.
- As we vary the error, the MLE recovers between 70 and 83% of the **efficiency** of the gold standard model.

Case Study

Piedmont Triad Food Access Landscape

Error-Prone Access

True Access

Low Access

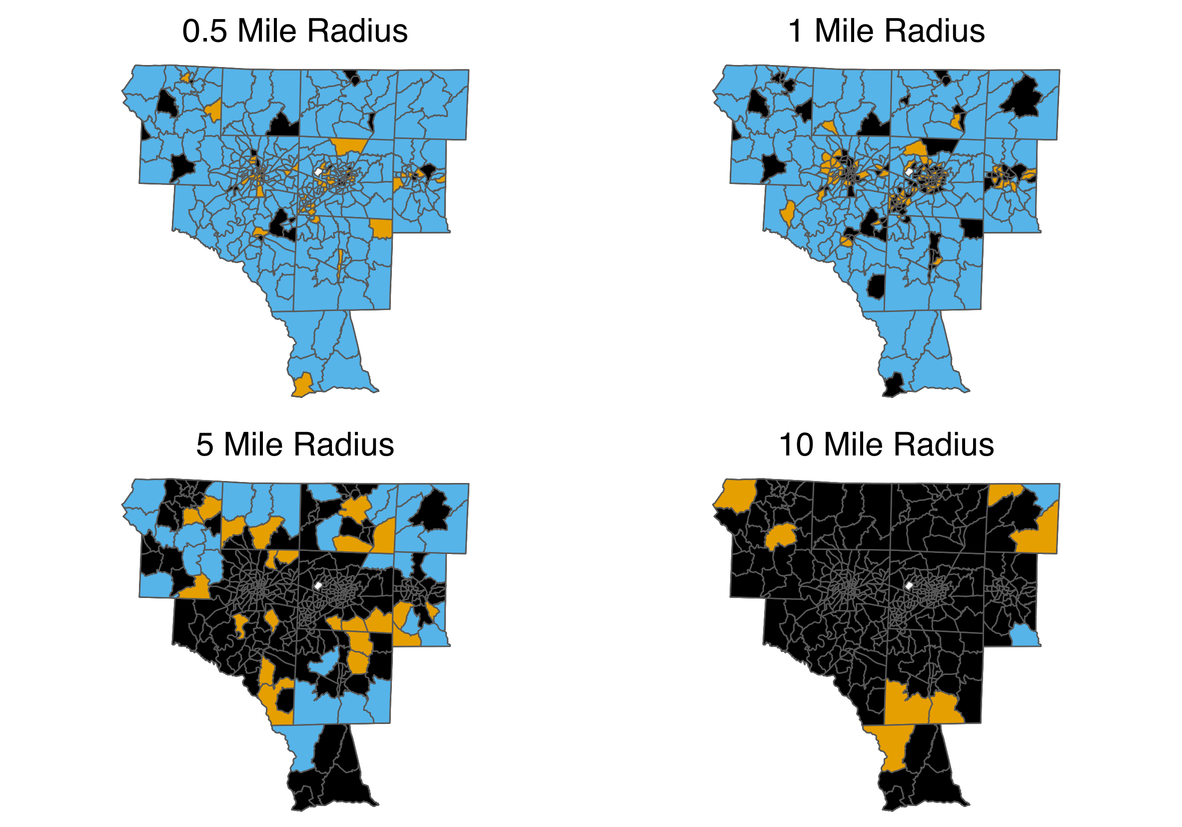

Figure: Food access landscape of the Piedmont triad

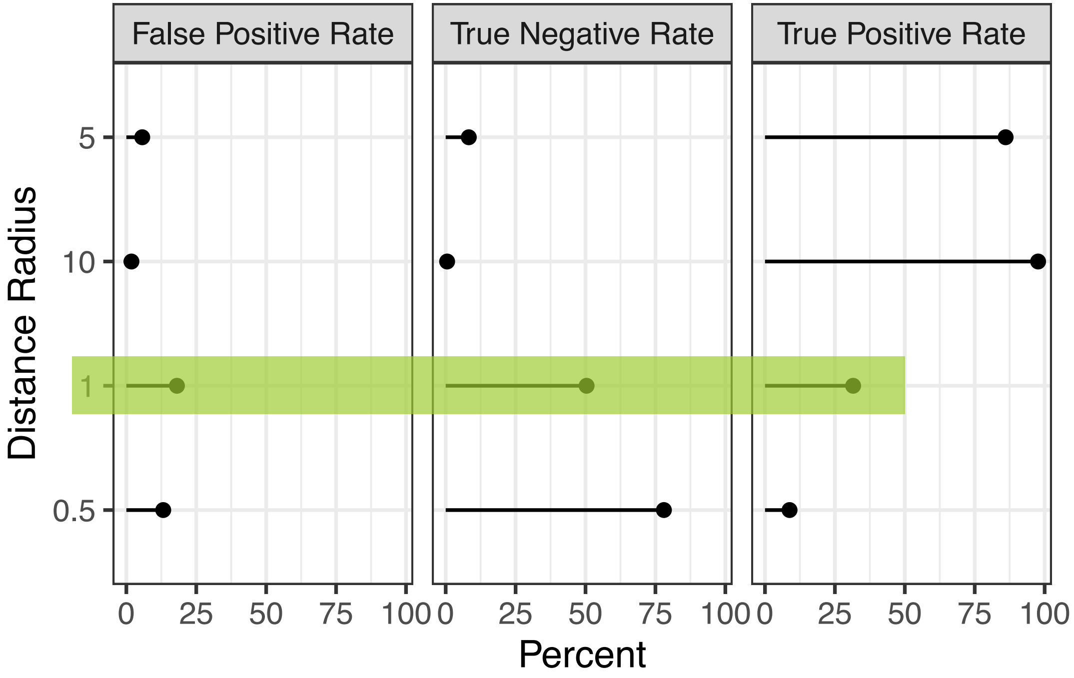

Figure: Summary of error rates in the Piedmont case study

Error Snapshot

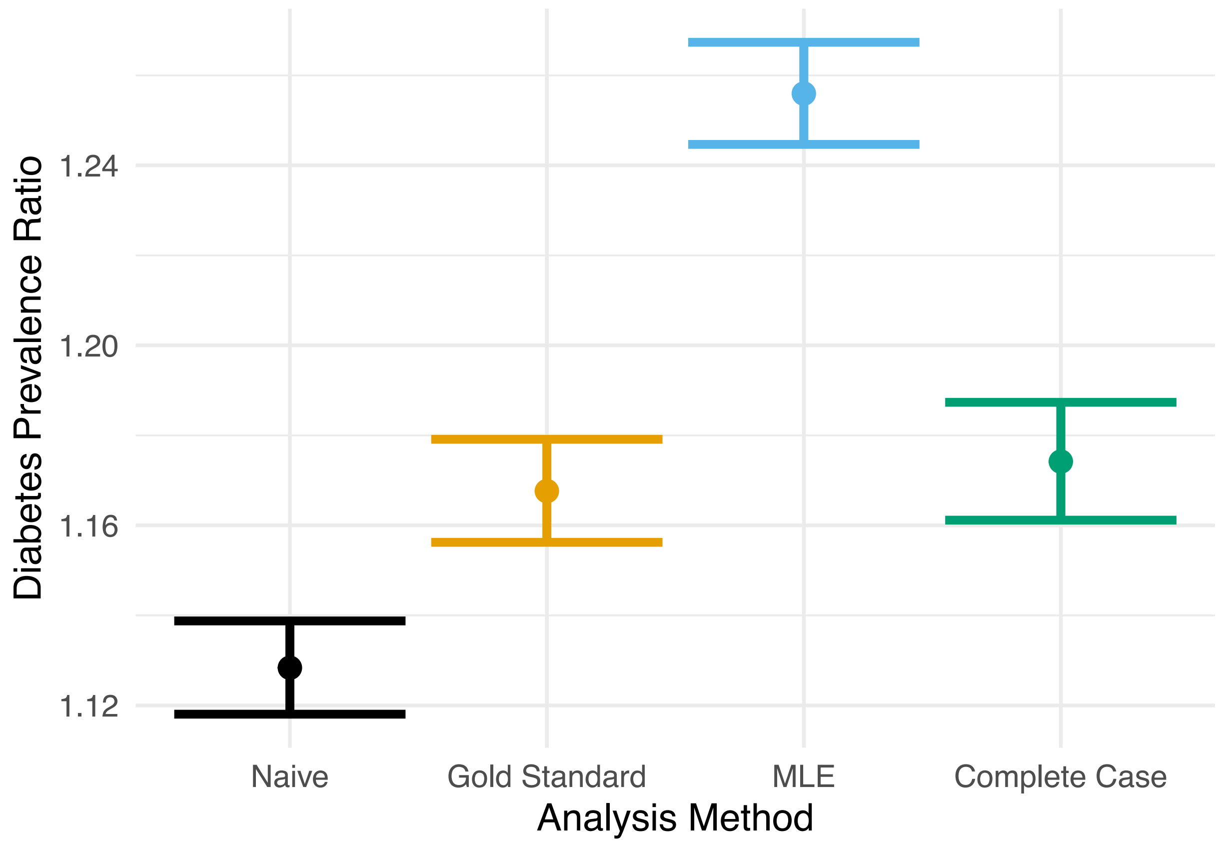

Figure: Diabetes prevalence estimates using four methods

One Mile Radius

Wrap-Up

Future Directions

- Expand case study
- Improve query design
- Tipping point analysis

References

- E. Gucciardi, M. Vahabi, N. Norris, J. P. Del Monte, and C. Farnum. The intersection between food insecurity and diabetes: a review. Current nutrition reports, 3:324–332, 2014.
- World Health Organization. Healthy diet, 2019. URL https://iris.who.int/handle/10665/325828.
- P. A. Shaw, R. H. Keogh, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 2—more complex methods of adjustment and advanced topics. Statistics in medicine, 39(16):2232–2263, 2020.
- L. Tang, R. H. Lyles, C. C. King, D. D. Celentano, and Y. Lo. Binary regression with differentially misclassified response and exposure variables. Statistics in Medicine, 34(9):1605–1620, 2015.
- S. Lotspeich, A. Mullan, L. D'Agostino McGowan, and S. Hepler. Combining straight-line and map-based distances to investigate food access and health. In Preparation, 2023+

THE ANDREW SABIN FAMILY

CENTER *for* ENVIRONMENT AND SUSTAINABILITY





Thank you!