CDC's National Center for Health Statistics

The Enhanced Modified Kalman Filter for Small Domain Estimates

Lauren M. Rossen[1] and Makram Talih[1,2]

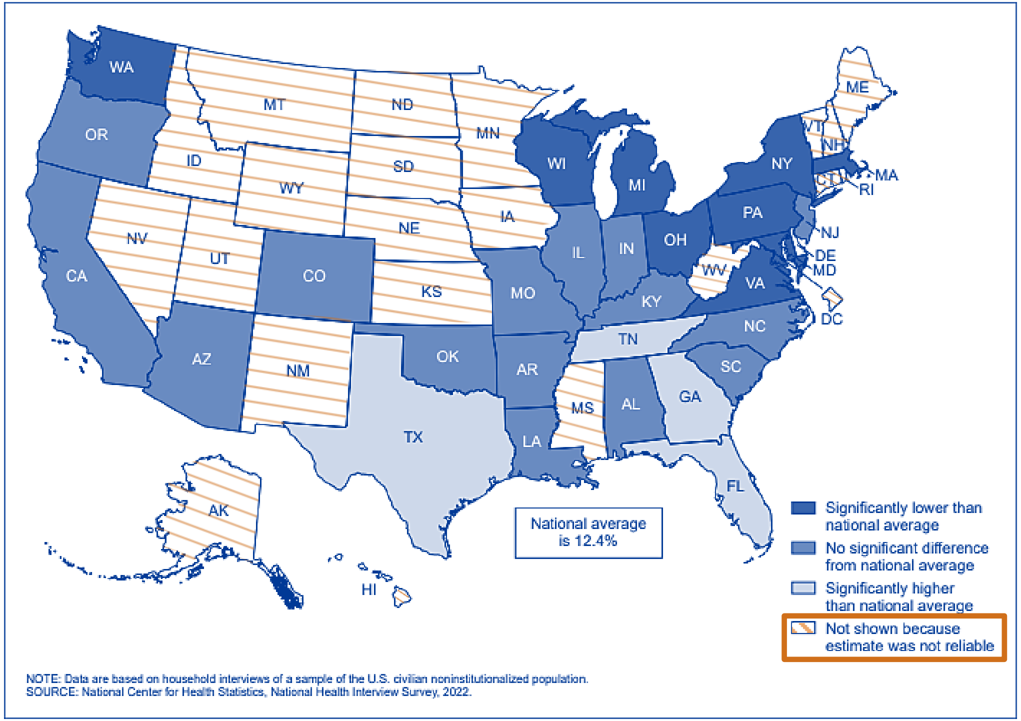
[1]Division of Research and Methodology
[2]Windsor Group, LLC

FCSM Research and Policy Conference, October 2024

Background

- Measuring health disparities for small subpopulations can be challenging because direct estimates may be unstable or unreliable due to small sample sizes
    - Estimates for these small subgroups are often suppressed

Figure 6. Adults ages 18–64 who were uninsured at time of interview: United States, 2022


WA
OR
CA
NV
ID
MT
WY
UT
AZ
NM
CO
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
OH
KY
TN
MS
AL
GA
FL
SC
NC
VA
WV
PA
NY
VT
NH
ME
MA
RI
CT
NJ
DE
MD
DC
AK
HI
Significantly lower than national average
No significant difference from national average
Significantly higher than national average
Not shown because estimate was not reliable

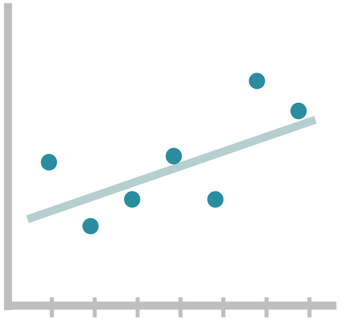
National average is 12.4%

NOTE: Data are based on household interviews of a sample of the U.S. civilian noninstitutionalized population.
SOURCE: National Center for Health Statistics, National Health Interview Survey, 2022.

Figure Source: Terlizzi EP, Cohen RA. Geographic variation in health insurance coverage: United States, 2022. National Health Statistics Reports; no 194. Hyattsville, MD: National Center for Health Statistics. 2023. DOI: https://doi.org/10.15620/ cdc:133320.

Background

- The Modified Kalman Filter (MKF) is a tool first released by RAND Corporation in 2011
  - Mixed effects models "borrowed strength" across time points and groups to predict estimates for smaller subpopulations
  - SAS macro that used pre-tabulated data as the input
    - Along with precompiled C code in an .exe file
  - Limited to equally-spaced time points and linear trends

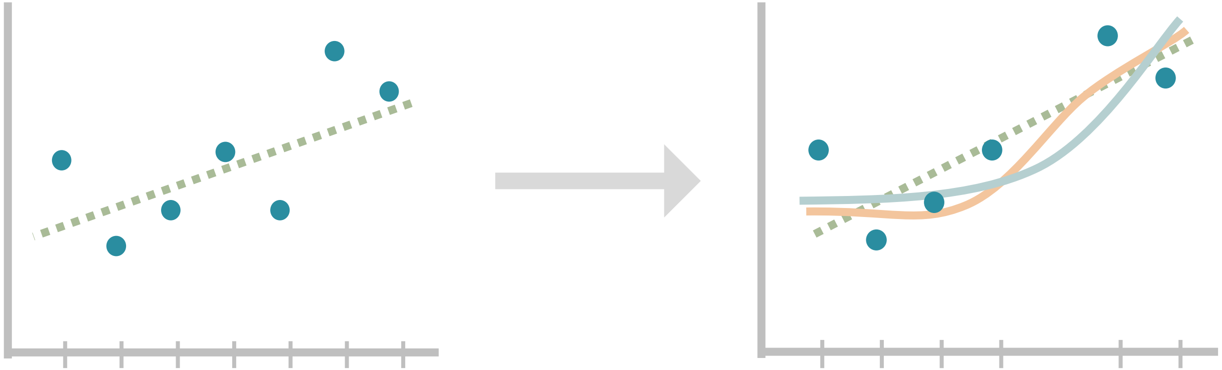


- We have developed an enhanced MKF macro to address some limitations and provide more functionality and transparency

Extensions/improvements to the MKF

The Enhanced MKF (eMKF) macro

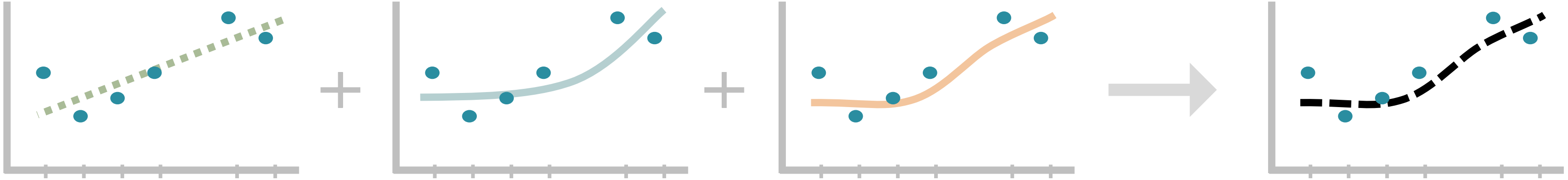
- Still uses mixed effects models to “borrow strength” across time points and groups to predict estimates for smaller subpopulations
- Allows for unequally spaced time points
  - Periodic content or survey/data collection disruptions
- Allows for non-linear trends (e.g., quadratic and cubic)



Extensions/improvements to the MKF

The Enhanced MKF (eMKF) macro

- Implements Bayesian model averaging
  - Better-fitting models given more weight
  - Guards against trend misspecification
  - Better accounts for uncertainty in trend model selection



- More flexible in terms of other model assumptions
  - Random sampling variances
- Easier to implement and interrogate (SAS macro)

What the eMKF macro does

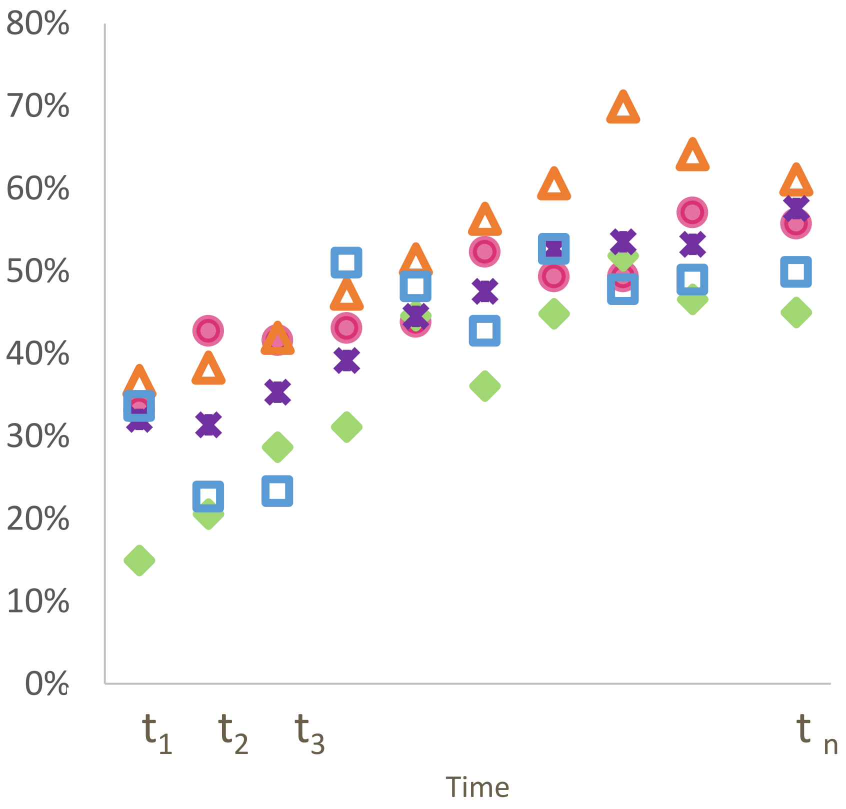


Figure for illustration purpose only

Mixed effects models

- Time trends (slopes) can be shared or differ by group

What the eMKF macro does

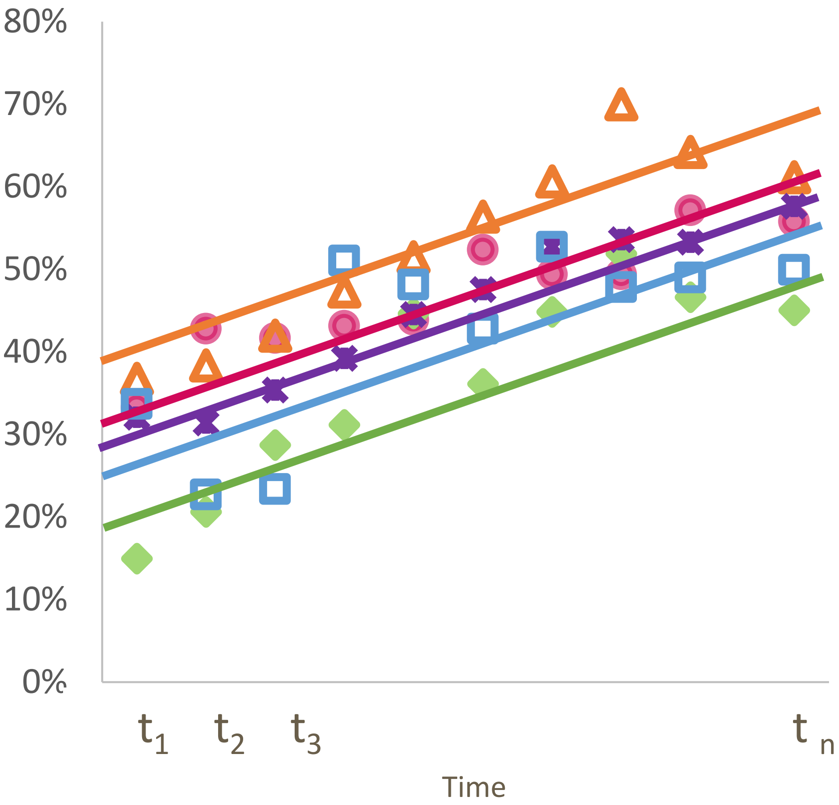


Figure for illustration purpose only

Mixed effects models

- Time trends (slopes) can be shared or differ by group
  - Common trends across groups = all groups have same slope

What the eMKF macro does

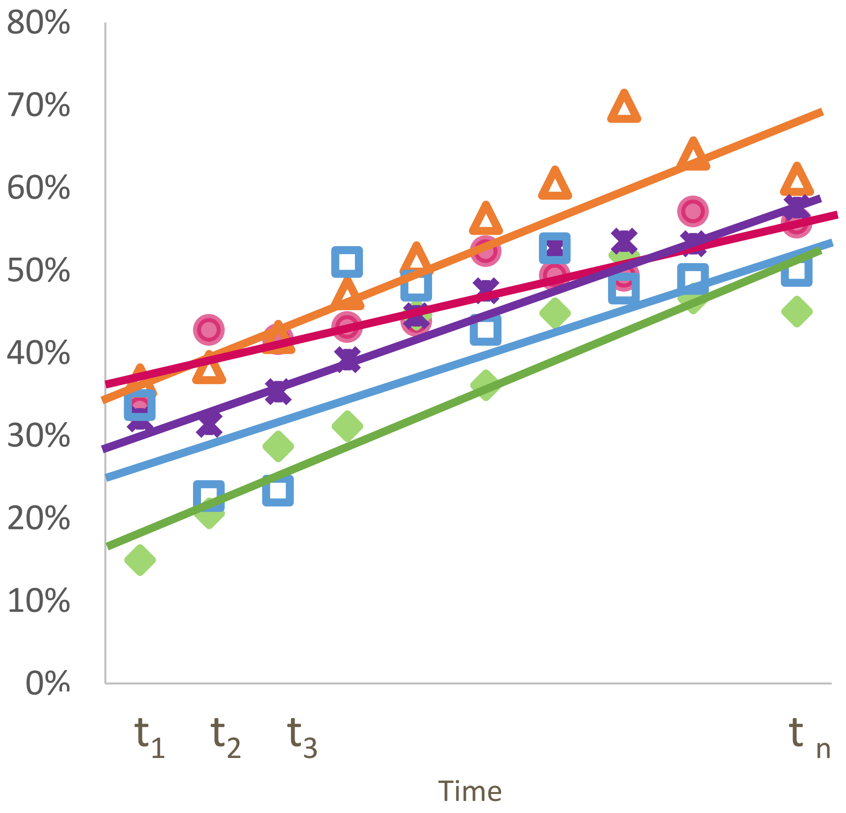


Figure for illustration purpose only

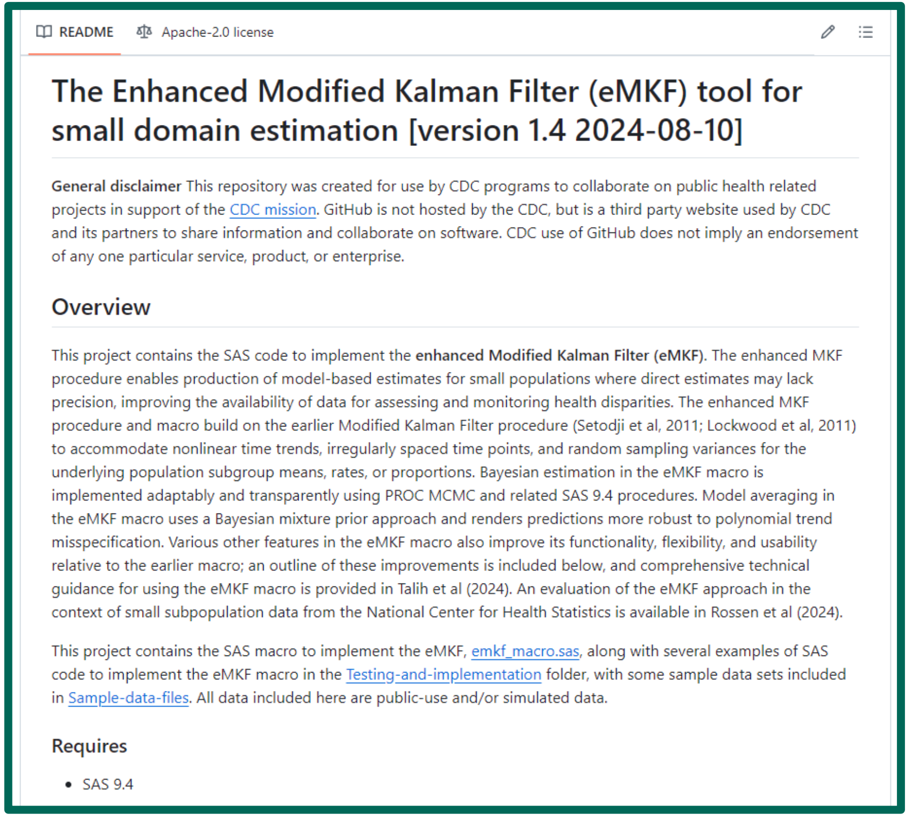
Mixed effects models

- Time trends (slopes) can be shared or differ by group
  - Common trends across groups = all groups have same slope
  - Independent trends across groups = slopes differ, but drawn from same distribution to ‘borrow strength’ over time and across groups
- Random (but auto-correlated) deviations from trend at times t_i

Documentation and Evaluation Reports

- Talih M, Rossen LM, Patel P, Earp M, Parker JD. Technical guidance for using the modified Kalman filter in small-domain estimation at the National Center for Health Statistics. National Center for Health Statistics. Vital Health Stat 2(209). 2024. DOI: https://dx.doi.org/10.15620/cdc/157496

- Rossen LM, Talih M, Patel P, Earp M, Parker JD. Evaluation of an enhanced modified Kalman filter approach for estimating health outcomes in small subpopulations. National Center for Health Statistics. Vital Health Stat 2(208). 2024. DOI: https://dx.doi.org/10.15620/cdc/157497



The Enhanced Modified Kalman Filter (eMKF) tool for small domain estimation [version 1.4 2024-08-10]

General disclaimer This repository was created for use by CDC programs to collaborate on public health related projects in support of the CDC mission. GitHub is not hosted by the CDC, but is a third party website used by CDC and its partners to share information and collaborate on software. CDC use of GitHub does not imply an endorsement of any one particular service, product, or enterprise.

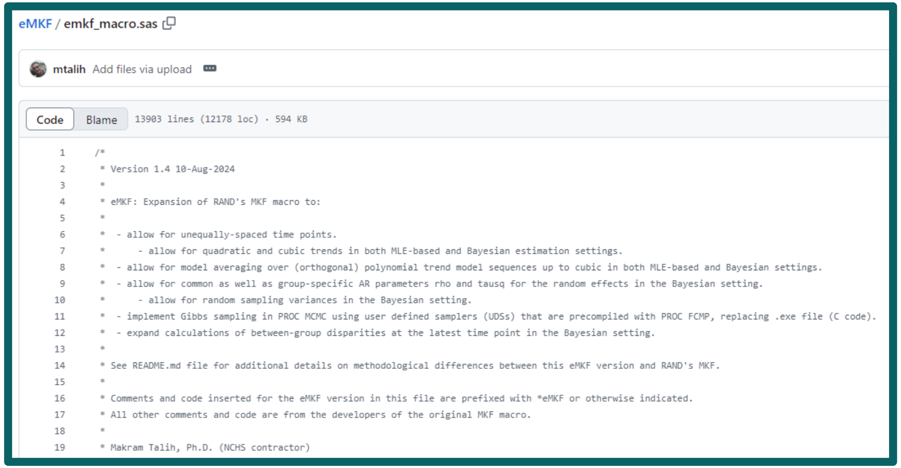
Overview

This project contains the SAS code to implement the **enhanced Modified Kalman Filter (eMKF)**. The enhanced MKF procedure enables production of model-based estimates for small populations where direct estimates may lack precision, improving the availability of data for assessing and monitoring health disparities. The enhanced MKF procedure and macro build on the earlier Modified Kalman Filter procedure (Setodji et al, 2011; Lockwood et al, 2011) to accommodate nonlinear time trends, irregularly spaced time points, and random sampling variances for the underlying population subgroup means, rates, or proportions. Bayesian estimation in the eMKF macro is implemented adaptably and transparently using PROC MCMC and related SAS 9.4 procedures. Model averaging in the eMKF macro uses a Bayesian mixture prior approach and renders predictions more robust to polynomial trend misspecification. Various other features in the eMKF macro also improve its functionality, flexibility, and usability relative to the earlier macro; an outline of these improvements is included below, and comprehensive technical guidance for using the eMKF macro is provided in Talih et al (2024). An evaluation of the eMKF approach in the context of small subpopulation data from the National Center for Health Statistics is available in Rossen et al (2024).

This project contains the SAS macro to implement the eMKF, emkf_macro.sas, along with several examples of SAS code to implement the eMKF macro in the Testing-and-implementation folder, with some sample data sets included in Sample-data-files. All data included here are public-use and/or simulated data.

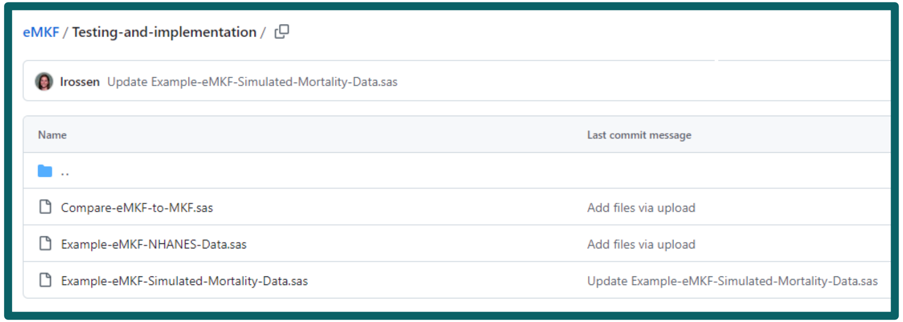
Requires

- SAS 9.4



13903 lines (12178 loc) · 594 KB

```
/*
 * Version 1.4 10-Aug-2024
 *
 * eMKF: Expansion of RAND's MKF macro to:
 *
 *  - allow for unequally-spaced time points.
 *     - allow for quadratic and cubic trends in both MLE-based and Bayesian estimation settings.
 *  - allow for model averaging over (orthogonal) polynomial trend model sequences up to cubic in both MLE-based and Bayesian settings.
 *  - allow for common as well as group-specific AR parameters rho and tausq for the random effects in the Bayesian setting.
 *     - allow for random sampling variances in the Bayesian setting.
 *  - implement Gibbs sampling in PROC MCMC using user defined samplers (UDSs) that are precompiled with PROC FCMP, replacing .exe file (C code).
 *  - expand calculations of between-group disparities at the latest time point in the Bayesian setting.
 *
 * See README.md file for additional details on methodological differences between this eMKF version and RAND's MKF.
 *
 * Comments and code inserted for the eMKF version in this file are prefixed with *eMKF or otherwise indicated.
 * All other comments and code are from the developers of the original MKF macro.
 *
 * Makram Talih, Ph.D. (NCHS contractor)
```


eMKF / Testing-and-implementation /
lrossen Update Example-eMKF-Simulated-Mortality-Data.sas

Name	Last commit message
..	
Compare-eMKF-to-MKF.sas	Add files via upload
Example-eMKF-NHANES-Data.sas	Add files via upload
Example-eMKF-Simulated-Mortality-Data.sas	Update Example-eMKF-Simulated-Mortality-Data.sas

https://github.com/CDCgov/eMKF

Evaluation of the eMKF

Simulated data

- National Health and Nutrition Examination Survey (NHANES; 1999 through March 2020)
  - Simulated trends in adult obesity (Body Mass Index [BMI] ≥ 30 kg/m^2)
  - By age and racial/ethnic group
  - Six simulated trend series: linear, quadratic, and cubic trends and trends that were common or group-specific to evaluate performance across trend shapes
- National Health Interview Survey (NHIS; 2019-2021)
  - Simulated trends (by year/quarter of interview) in diagnosed diabetes
  - By age and racial/ethnic group
  - Three simulated trend series based on small (10%), medium (20%) and large (40%) subsamples of the total NHIS to evaluate performance by sample size

Analysis

- Used the eMKF Bayesian Model Average
  - Combines estimates from 7 models
  - No trend, linear, quadratic, and cubic trends
  - Trends shared across groups or different for each group

- Model performance/accuracy assessed using the root mean squared error (RMSE)

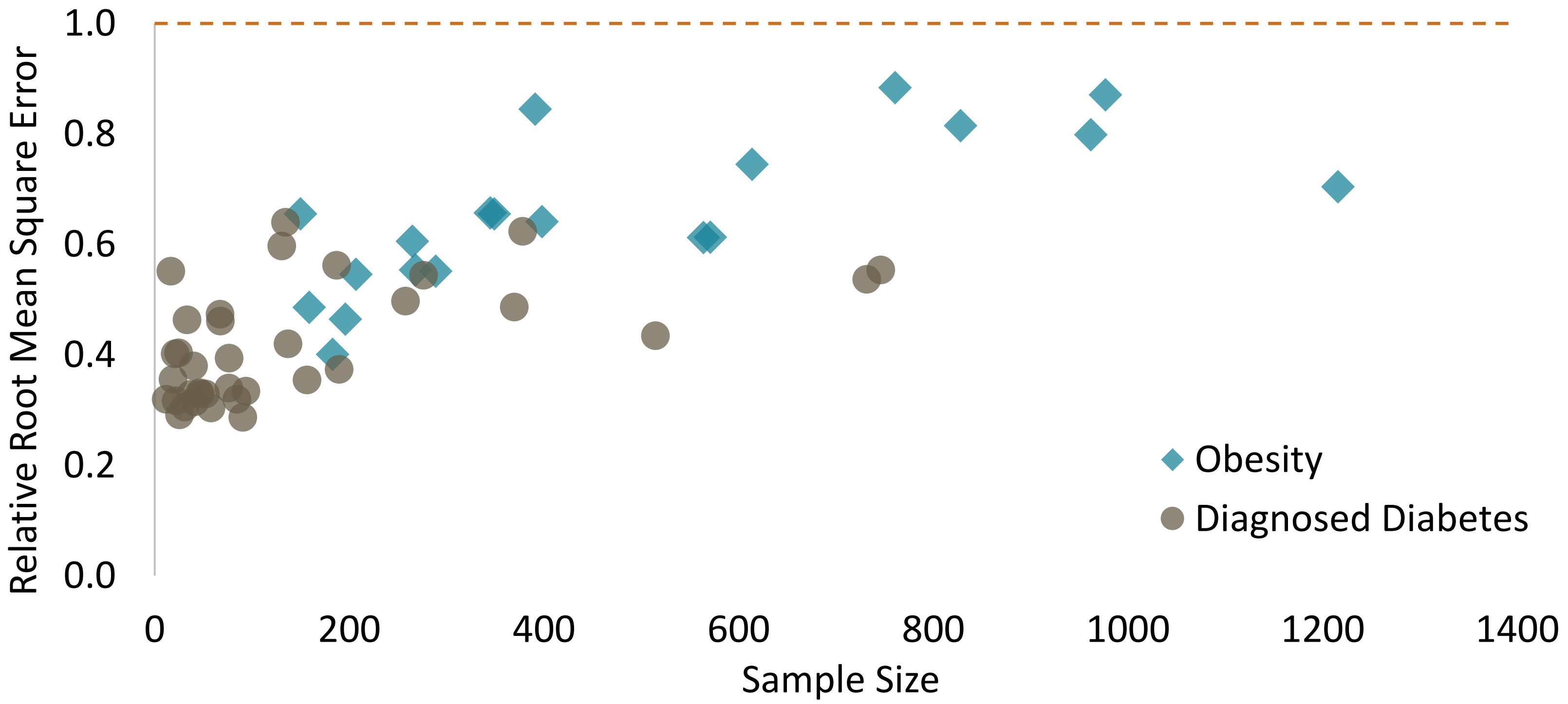
$$RMSE = \sqrt{standard\ error^2 + bias^2}$$

- Relative RMSE <1 indicates improved accuracy/precision of the eMKF estimates compared with the direct estimates

How accurate are the eMKF estimates?

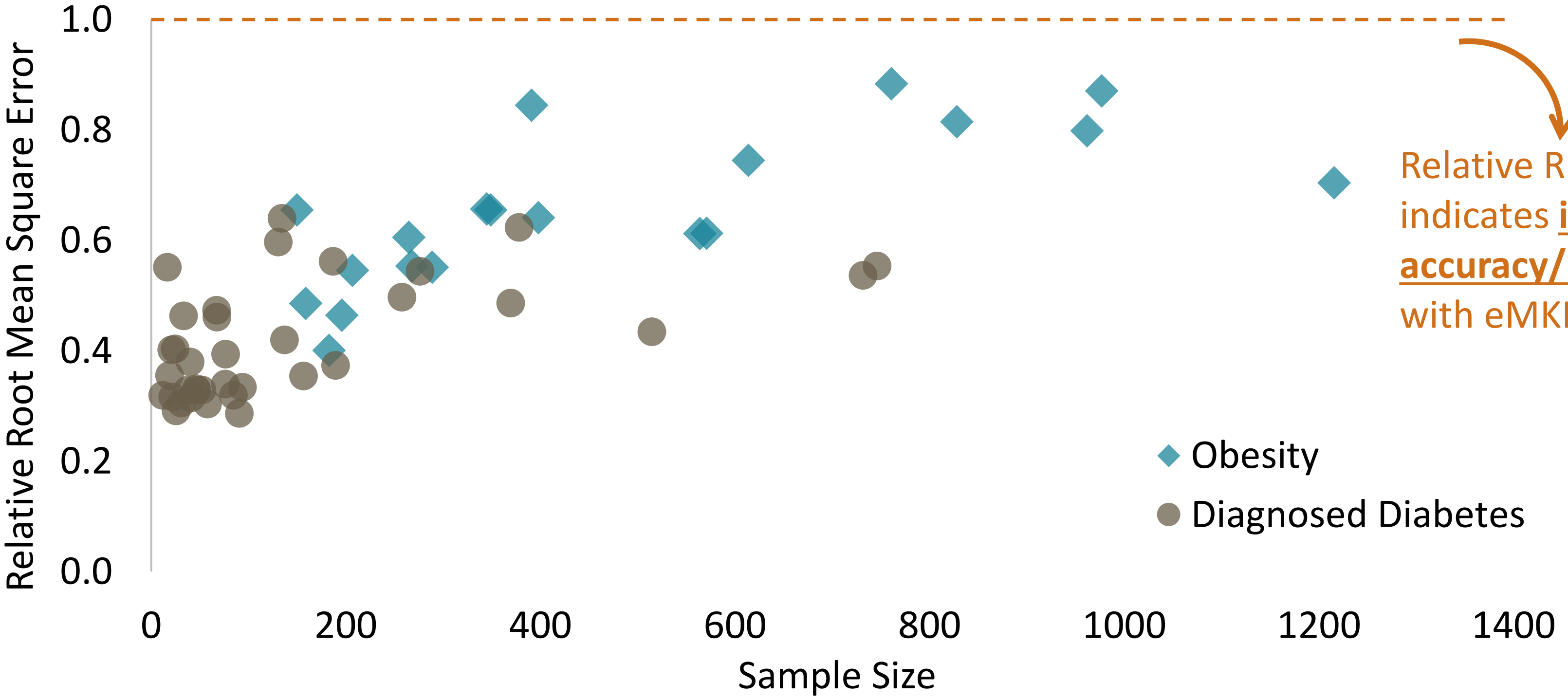
Relative accuracy of eMKF estimates of diagnosed diabetes and obesity compared with direct estimates

Simulated NHIS and NHANES data



Relative accuracy of eMKF estimates of diagnosed diabetes and obesity compared with direct estimates

Simulated NHIS and NHANES data

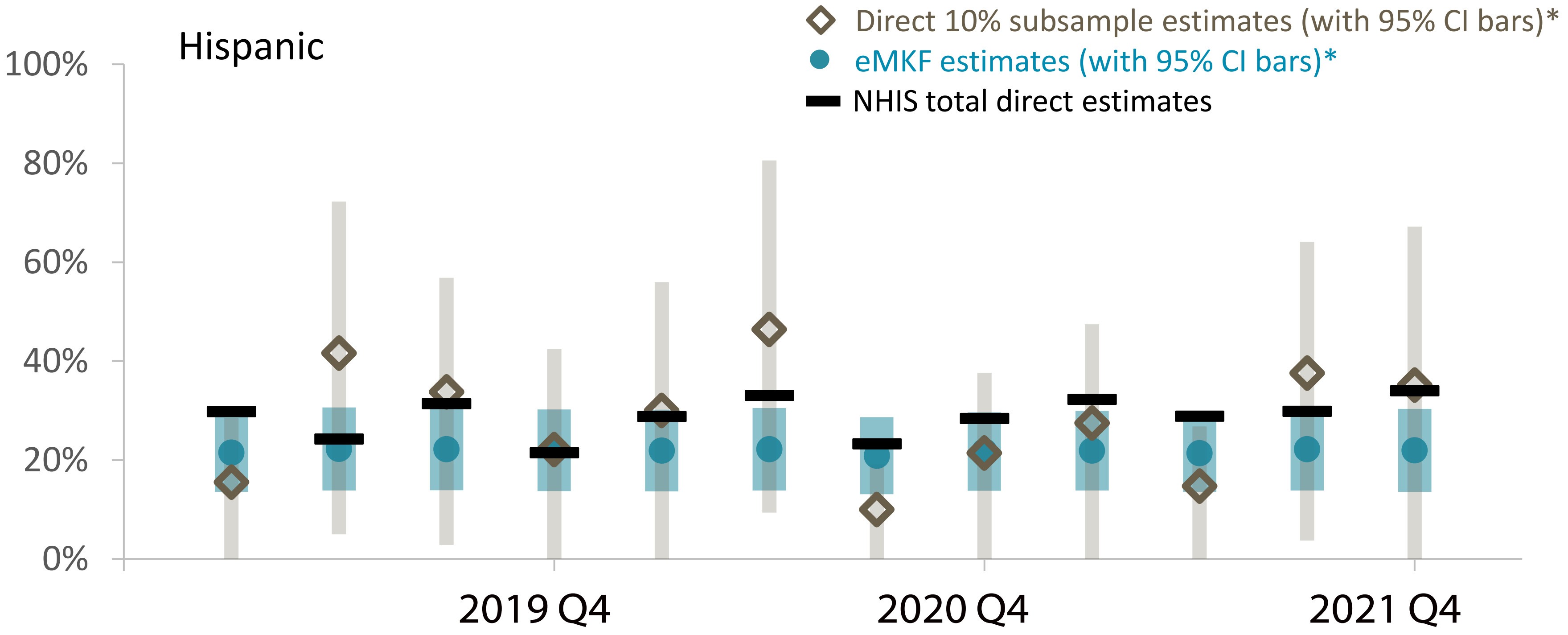


Relative RMSE <1 indicates **improved accuracy/precision** with eMKF

Direct and model-based estimates of prevalence of diagnosed diabetes

Adults age 65 and older, by race/ethnicity and quarter of interview

10% subsample of NHIS data

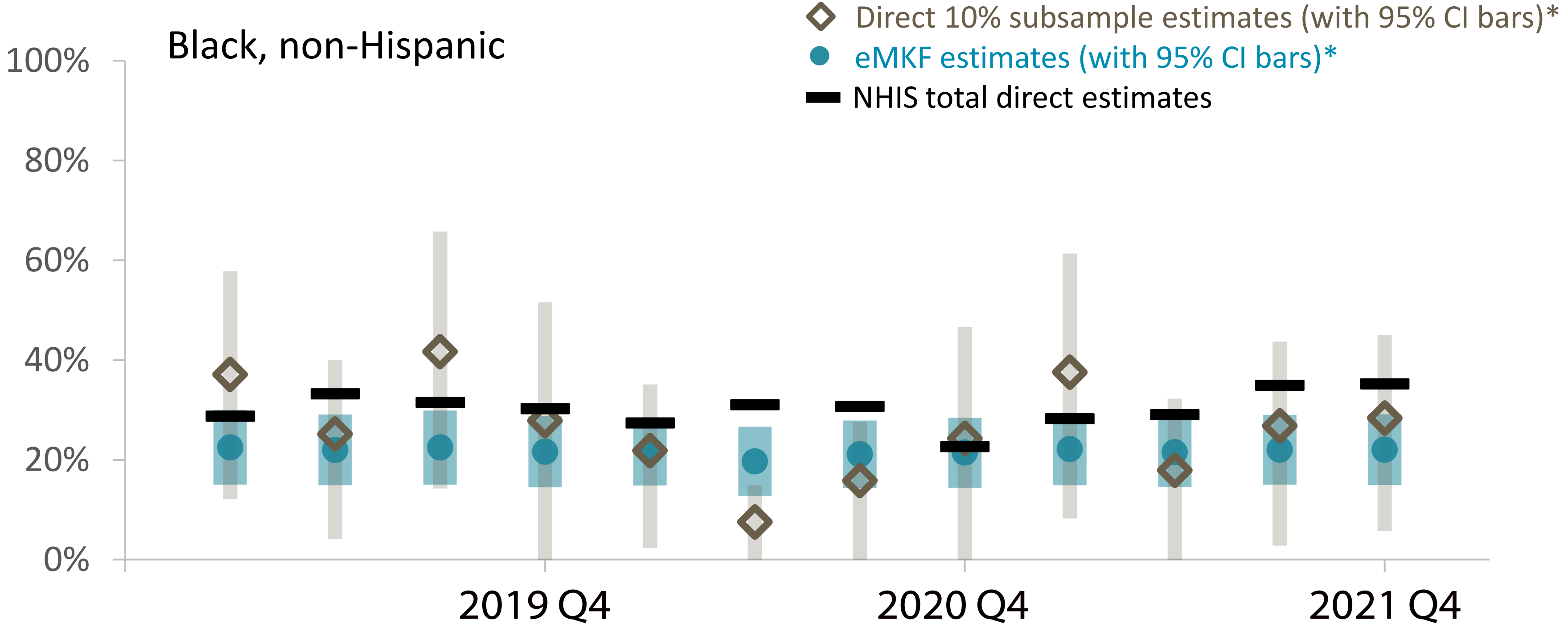


**Estimates are based on subsample sizes of 11-19 in each quarter*

Direct and model-based estimates of prevalence of diagnosed diabetes

Adults age 65 and older, by race/ethnicity and quarter of interview

10% subsample of NHIS data

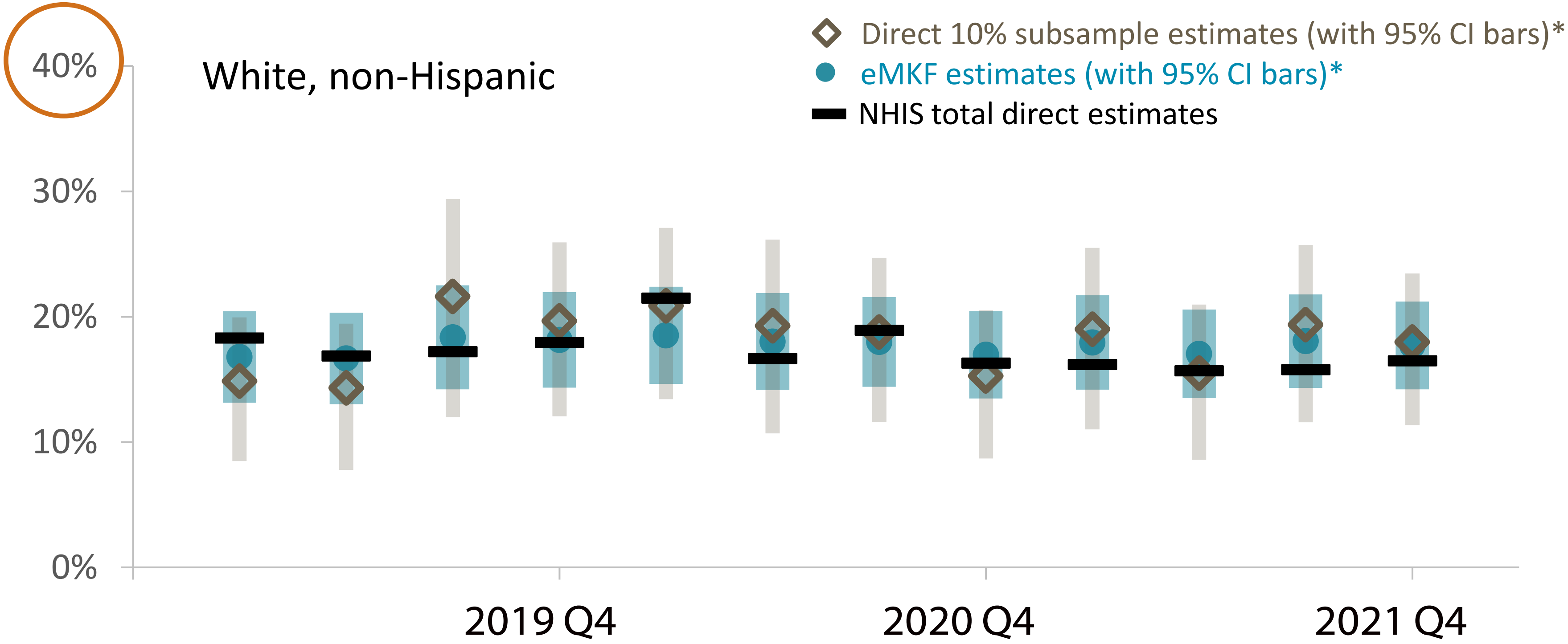


*Estimates are based on subsample sizes of 13-25 in each quarter

Direct and model-based estimates of prevalence of diagnosed diabetes

Adults age 65 and older, by race/ethnicity and quarter of interview

10% subsample of NHIS data

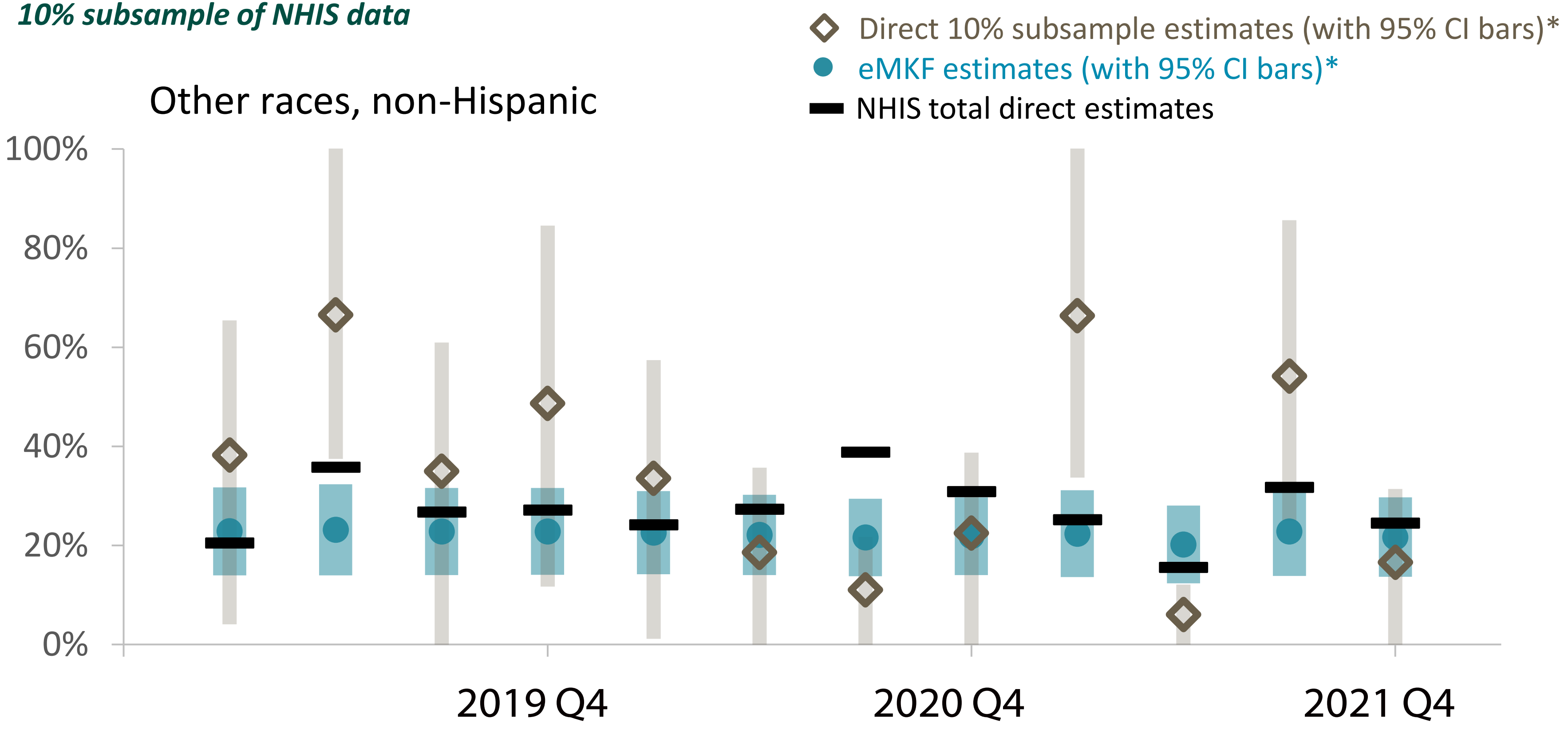


**Estimates are based on subsample sizes of 150-240 in each quarter*

Direct and model-based estimates of prevalence of diagnosed diabetes

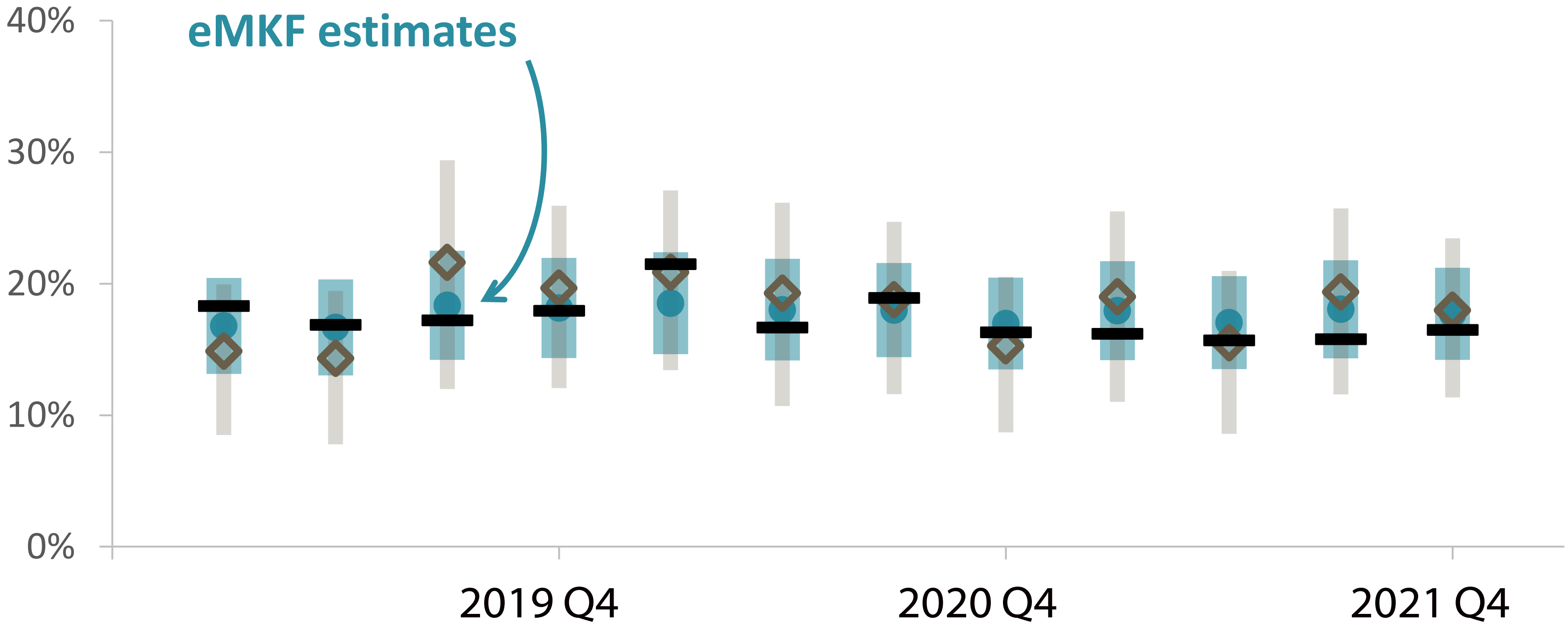
Adults age 65 and older, by race/ethnicity and quarter of interview

10% subsample of NHIS data



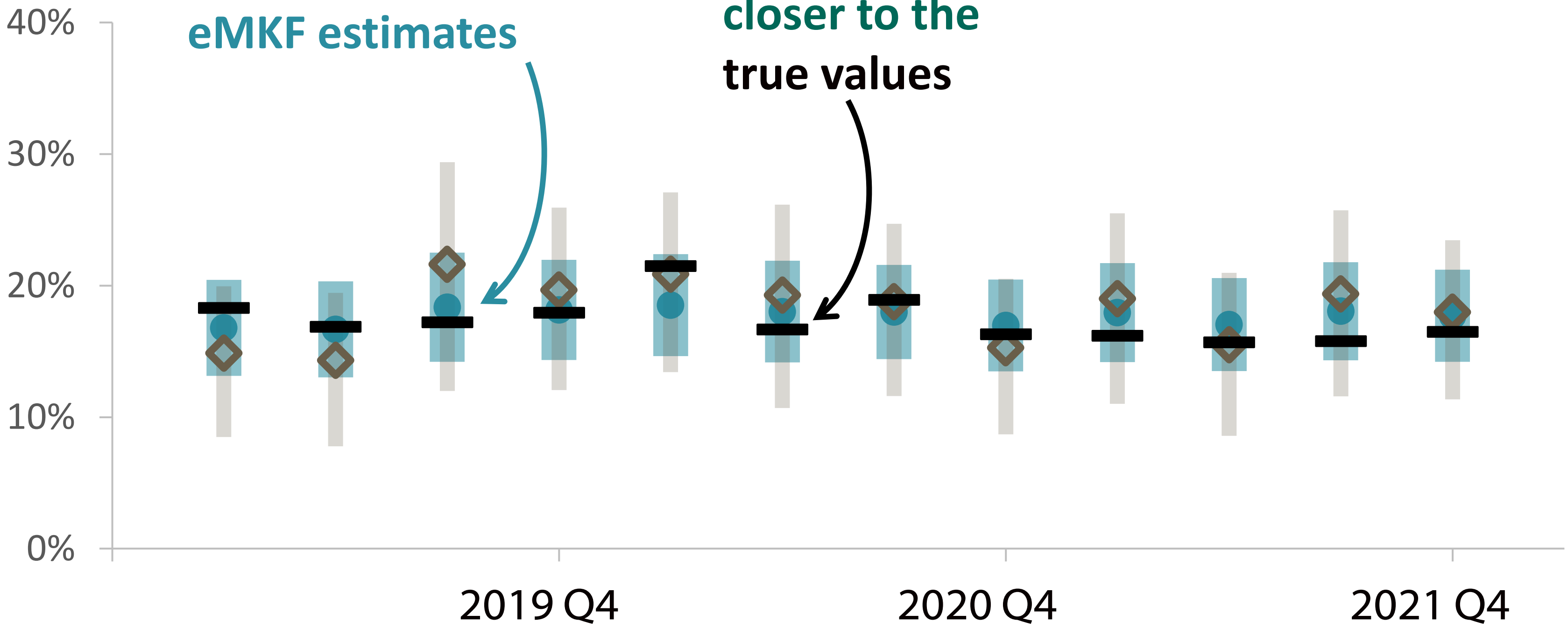
**Estimates are based on sample sizes of 10-13 in each quarter*

For the largest group where we have stable estimates, the eMKF estimates



For the largest group where we have stable estimates, the eMKF estimates

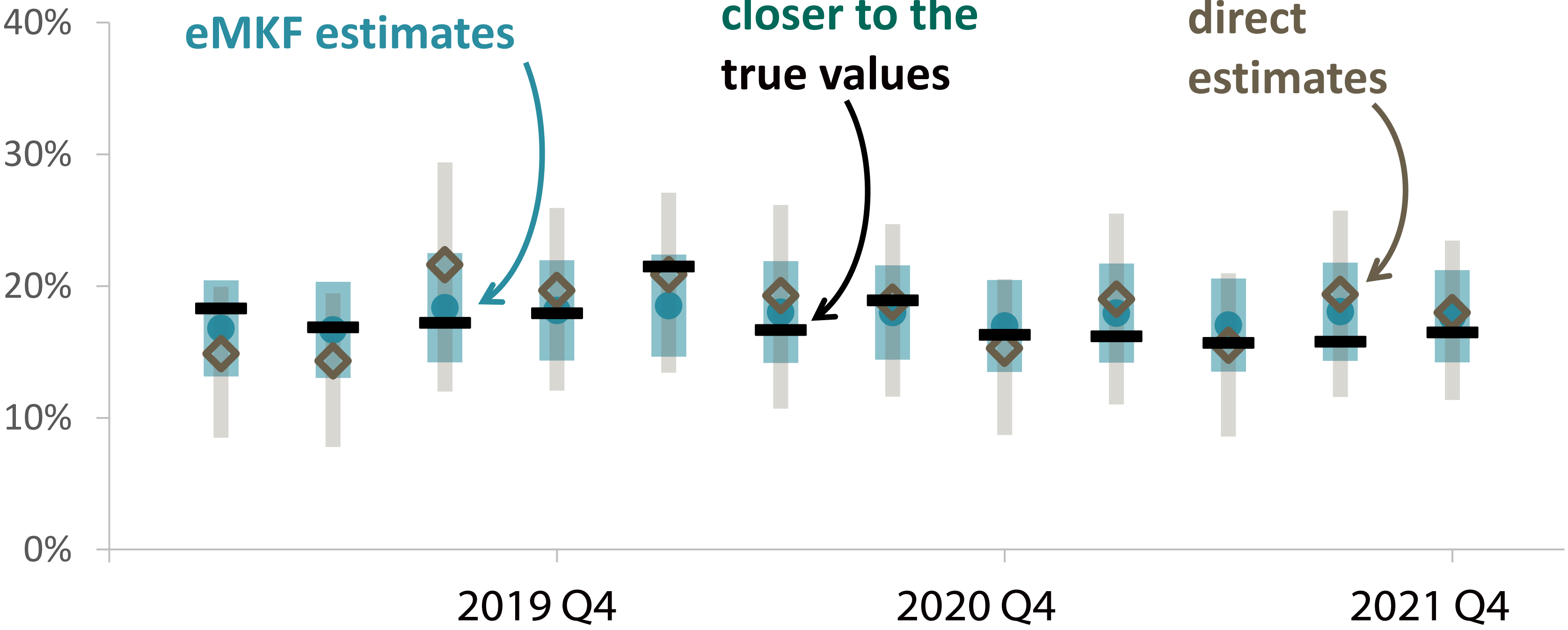
are typically closer to the true values



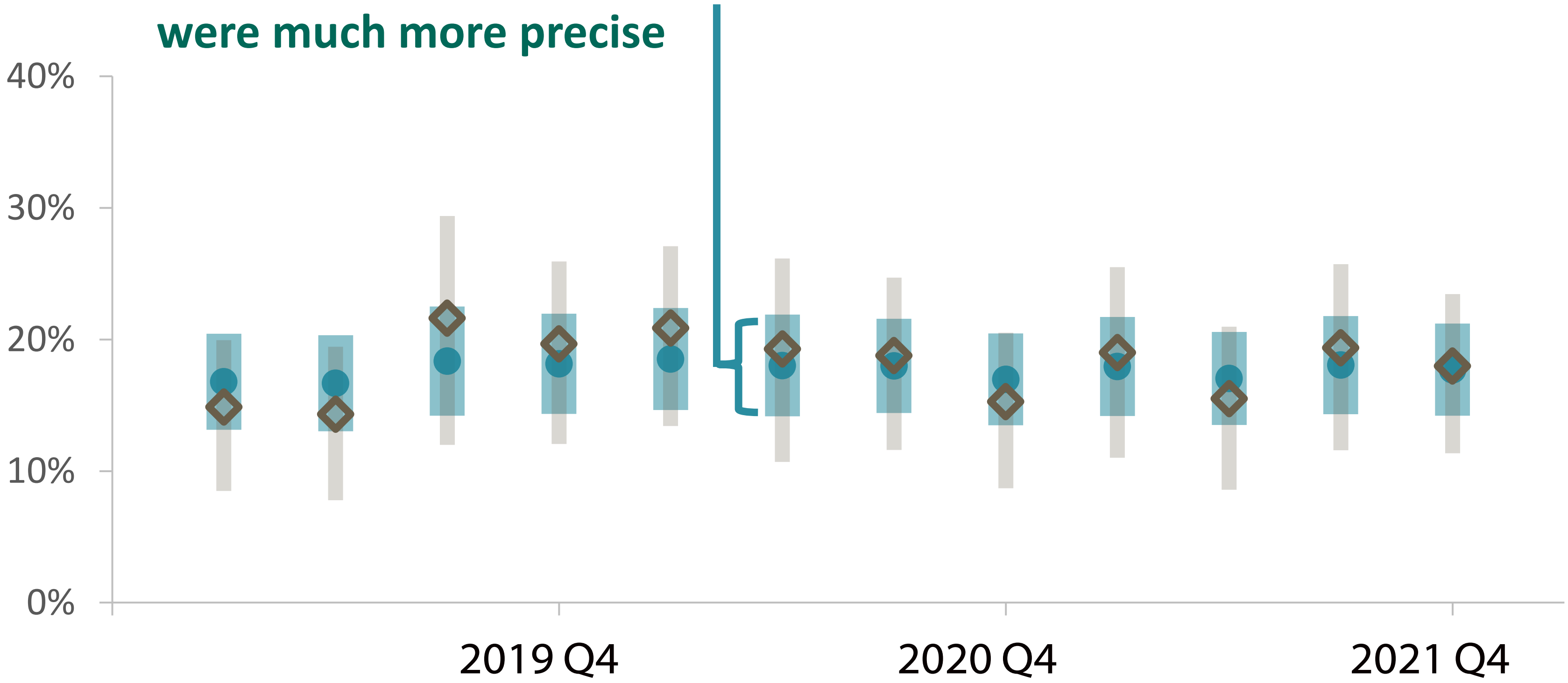
For the largest group where we have stable estimates, the eMKF estimates

are typically closer to the true values

than the subsample direct estimates

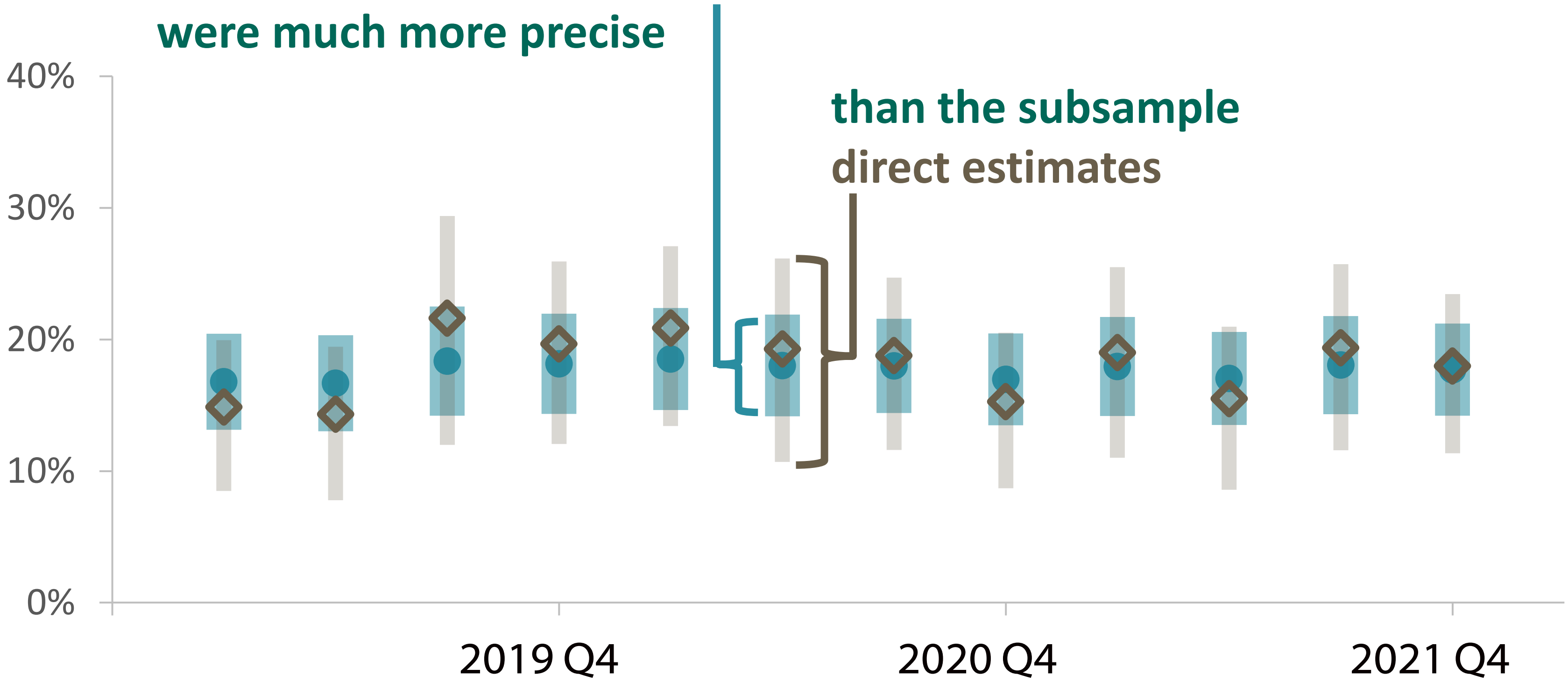


Even for the largest group where we have stable estimates, the eMKF estimates were much more precise

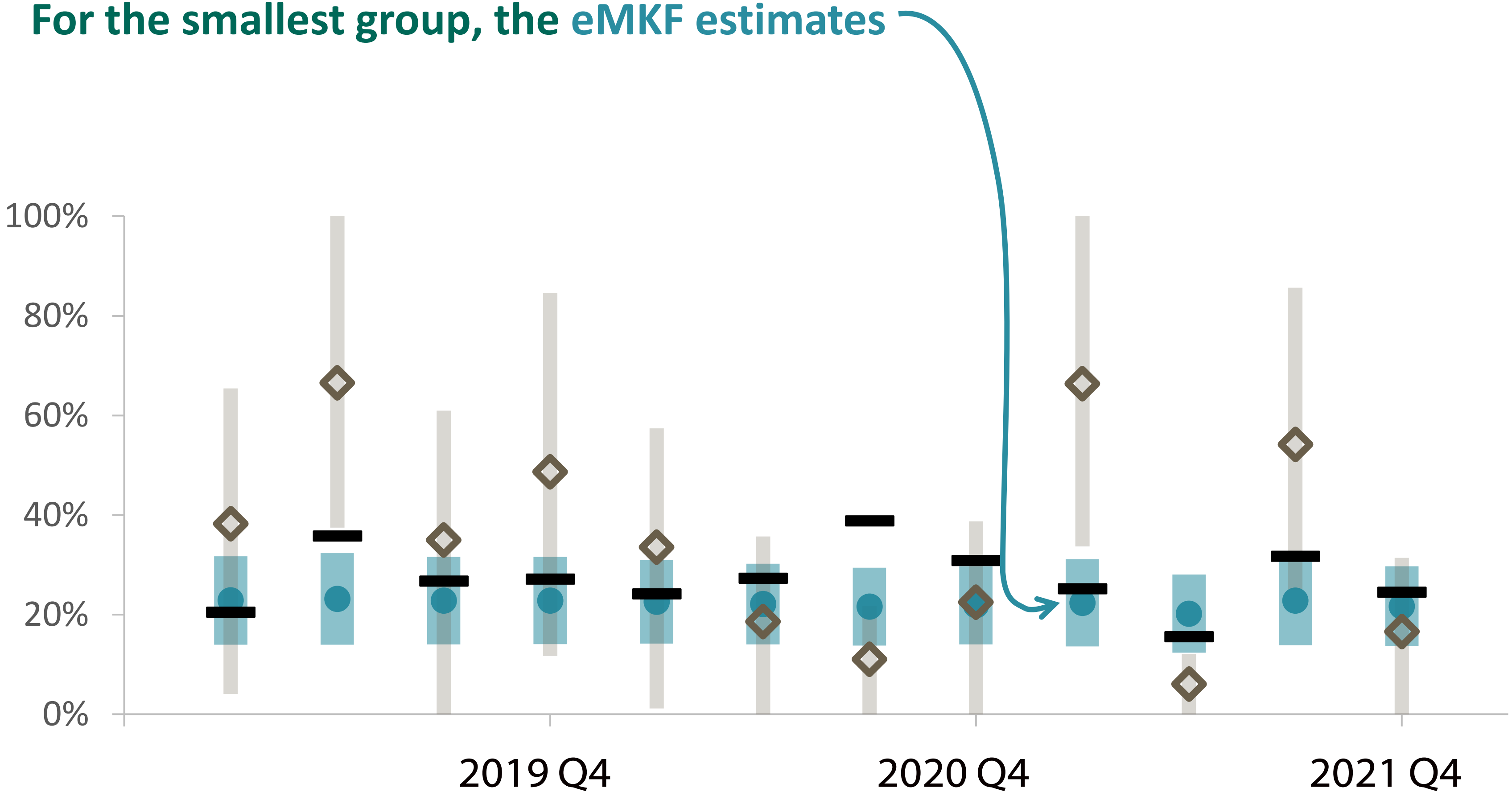


Even for the largest group where we have stable estimates, the eMKF estimates were much more precise

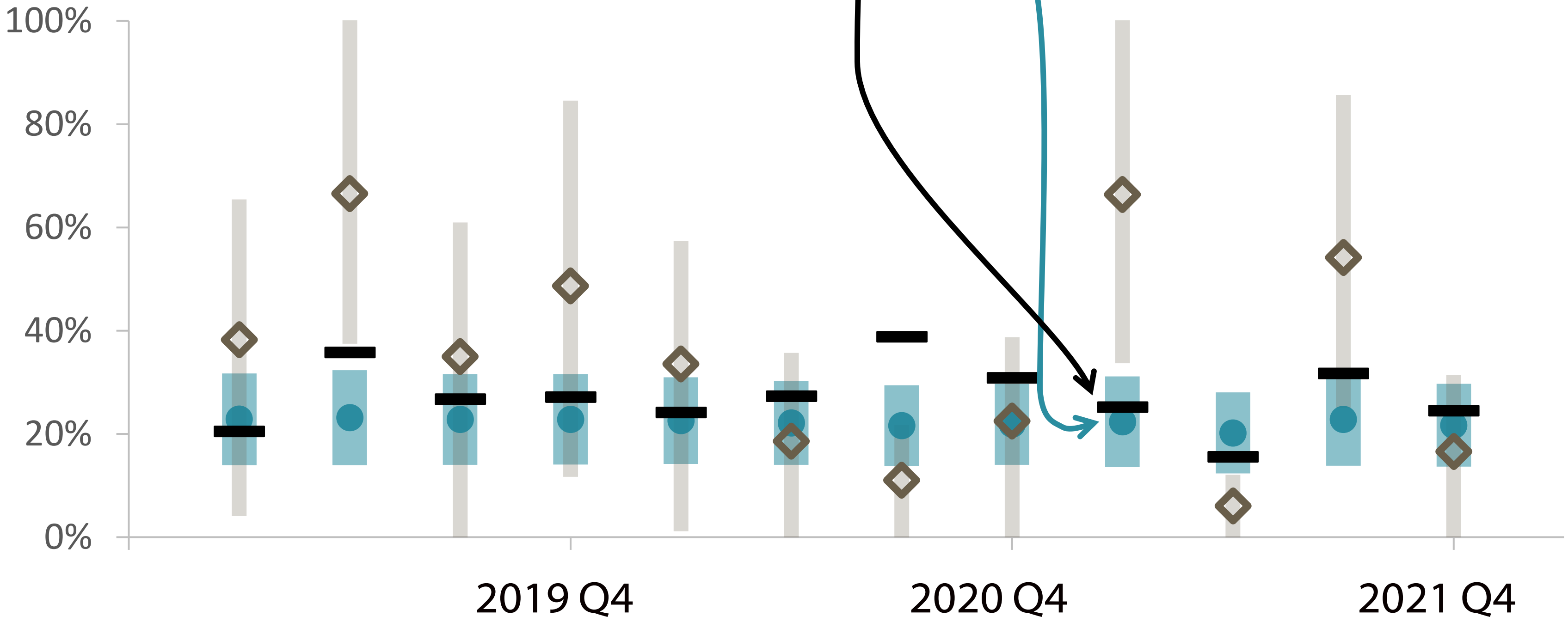
than the subsample direct estimates



For the smallest group, the eMKF estimates

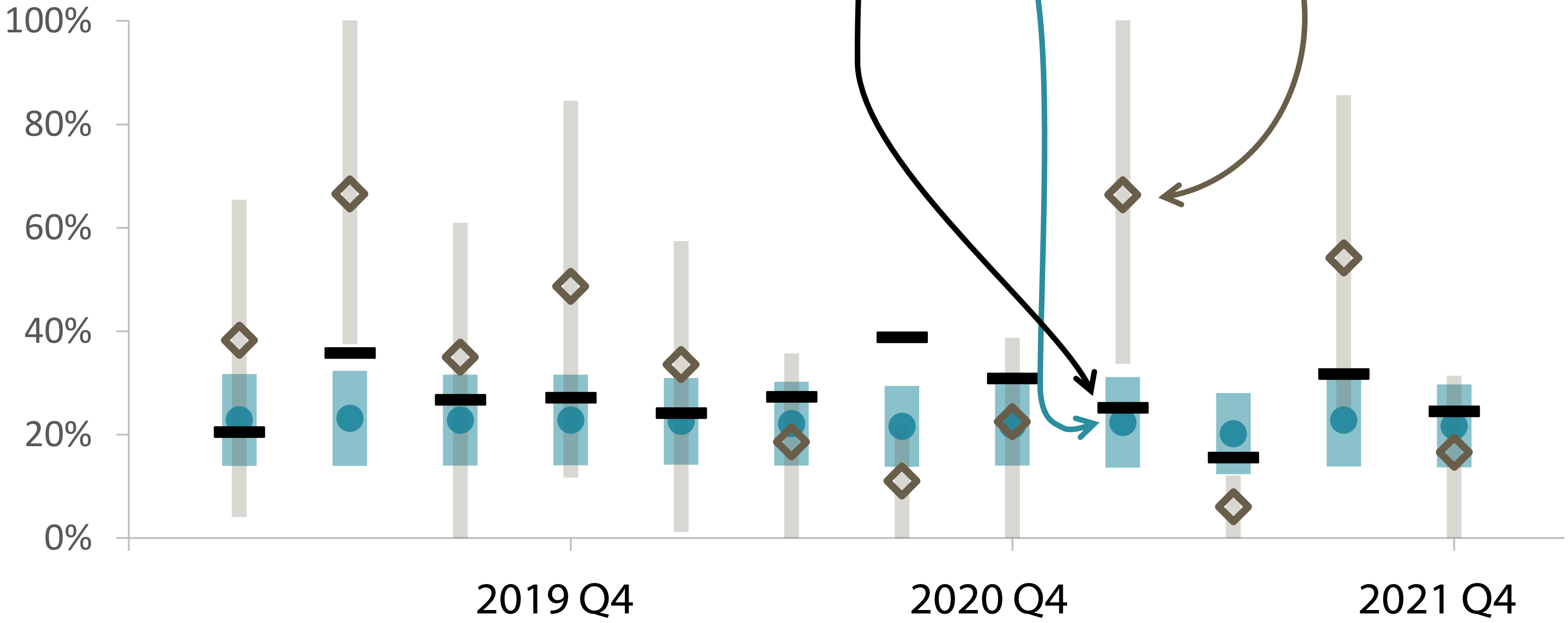


For the smallest group, the eMKF estimates were also generally closer to the true values



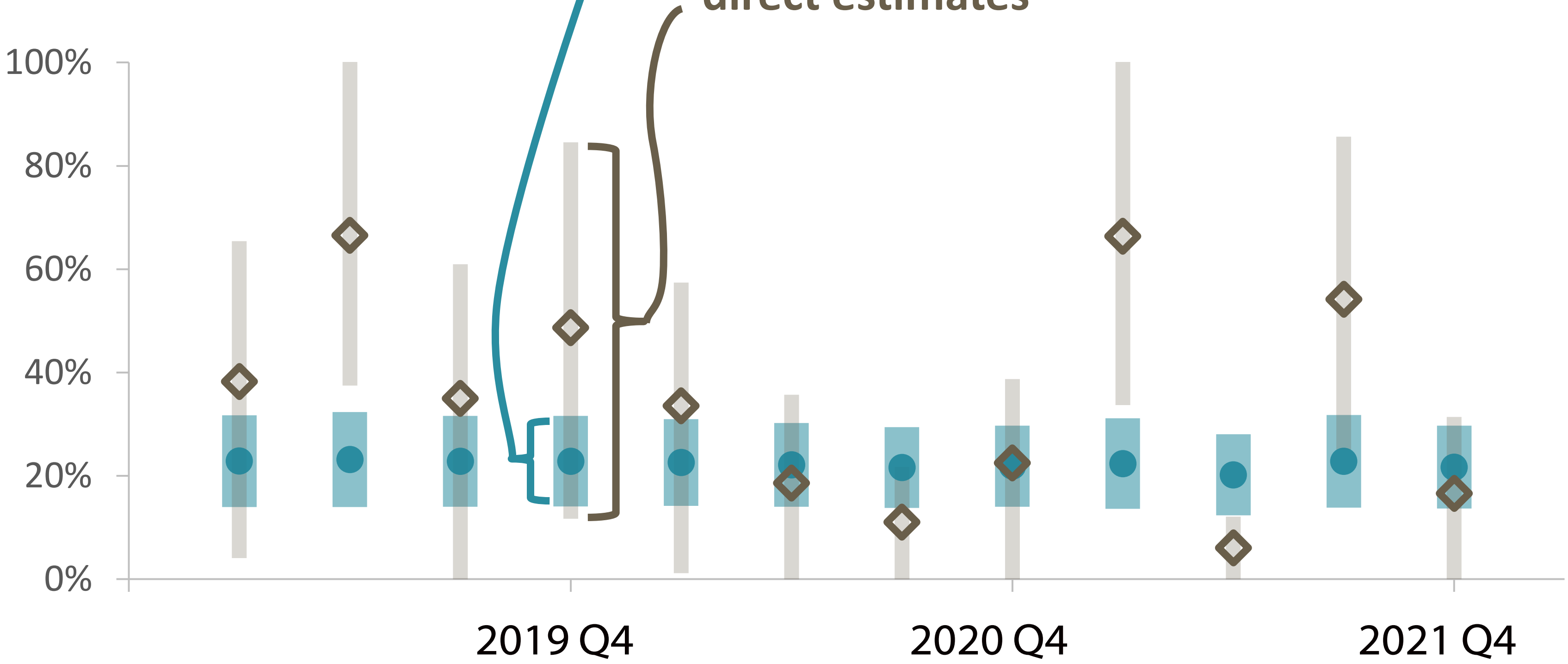
For the smallest group, the eMKF estimates were also generally closer to the true values

than the subsample direct estimates



And the eMKF estimates were much more precise

than the subsample direct estimates

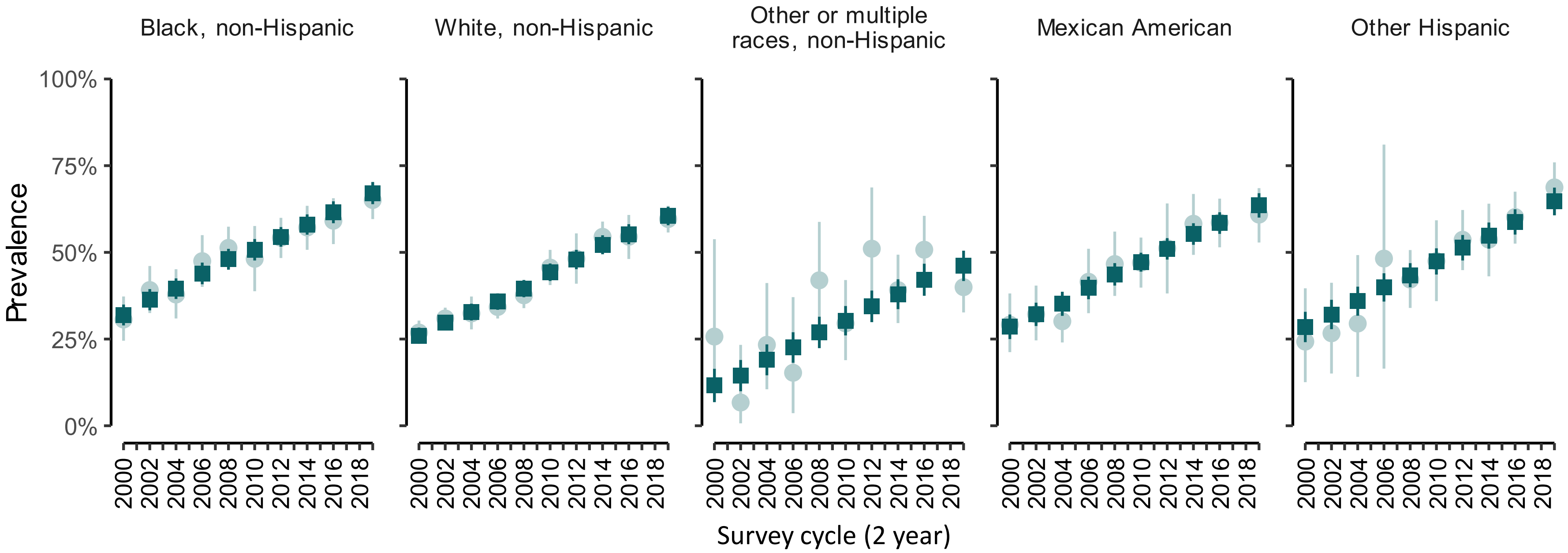


How well does the BMA eMKF option capture various trends?

Direct and model-based estimates of prevalence of obesity

Adults age 65 and older, by race/ethnicity and year

Simulated NHANES data where groups had a common ***linear*** *trend*

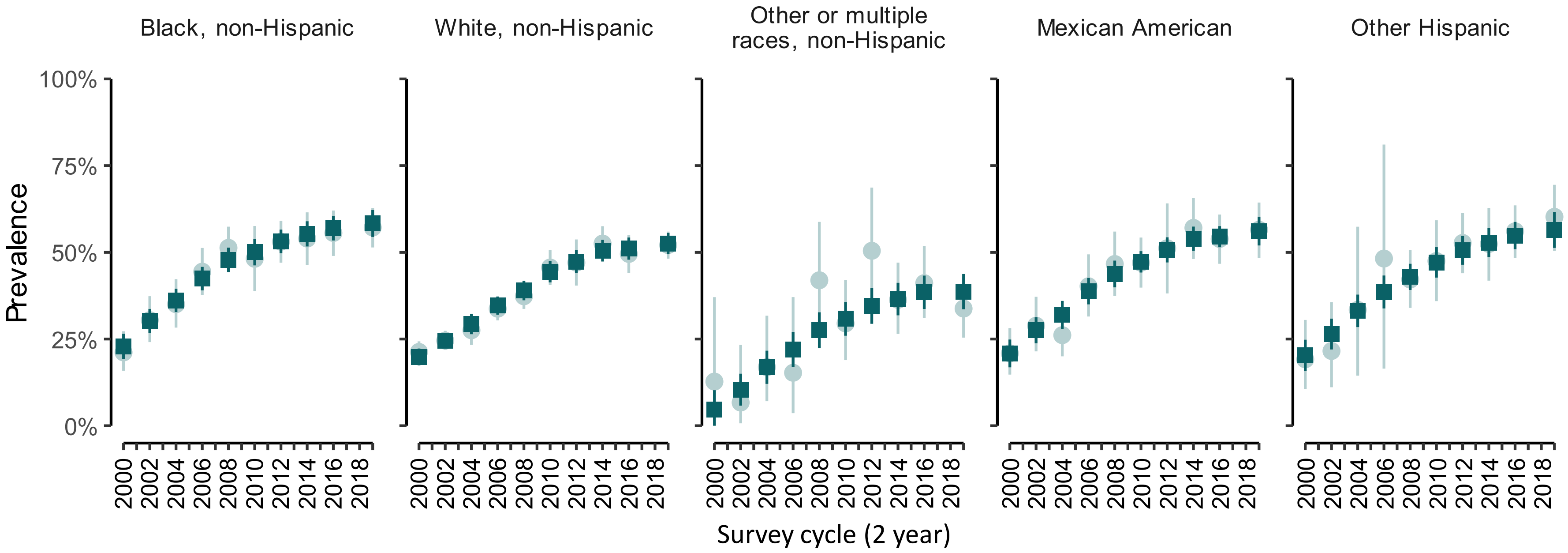


■ **eMKF estimates (with 95% CI bars)** ● **Direct estimates (with 95% CI bars)**

Direct and model-based estimates of prevalence of obesity

Adults age 65 and older, by race/ethnicity and year

Simulated NHANES data where groups had a common ***quadratic*** *trend*

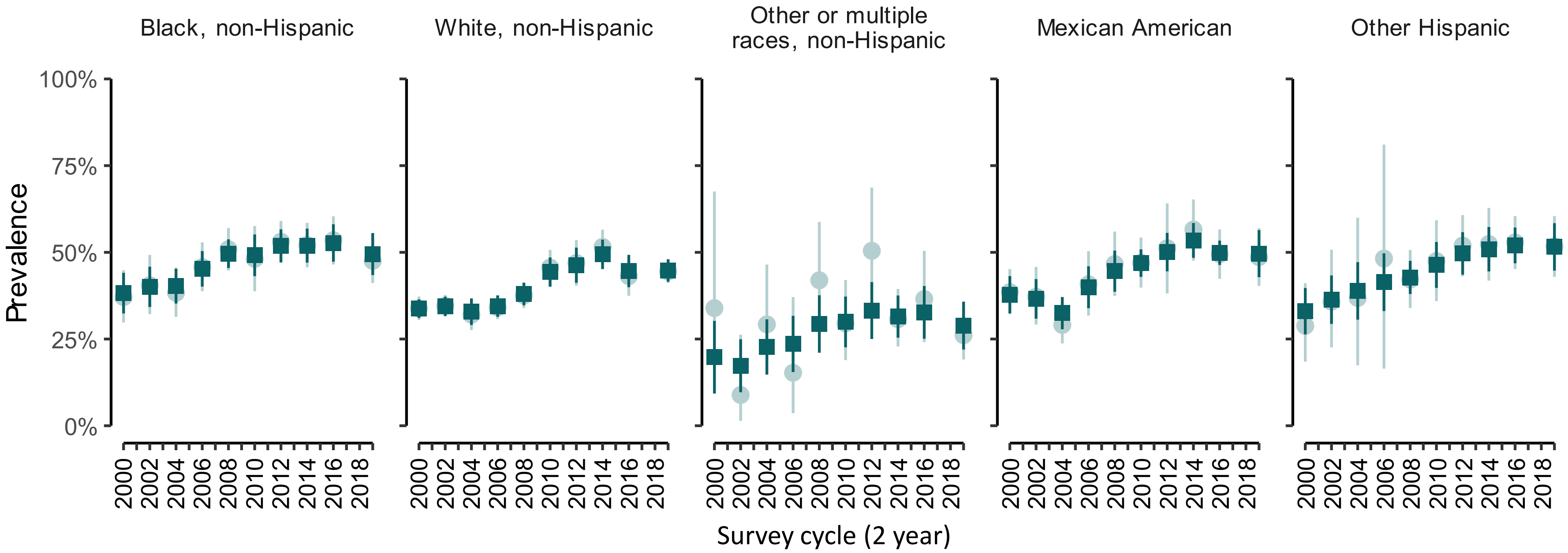


■ **eMKF estimates (with 95% CI bars)** ● **Direct estimates (with 95% CI bars)**

Direct and model-based estimates of prevalence of obesity

Adults age 65 and older, by race/ethnicity and year

*Simulated NHANES data where groups had a common **cubic** trend*



■ **eMKF estimates (with 95% CI bars)** ● **Direct estimates (with 95% CI bars)**

Summary

- The eMKF tool resulted in marked improvements in RMSE relative to direct estimates, with larger improvements for smaller sample sizes
- Improvements were seen across a wide array of simulated analytic scenarios
- In all cases, relative RMSEs were smaller for model-based estimates than direct estimates
- Gains in equivalent sample size (1/relative RMSE) of up to 420%

Limitations

- Requires ≥k+4 time points to fit a degree k polynomial trend
  - Linear trend ➔ need 5 time points
  - Quadratic ➔ need 6 time points
  - Cubic ➔ need 7 time points
- Requires no missing/non-sampled time periods for any group
  - Need to aggregate to larger units if there are missing/non-sampled groups
  - Need to provide direct estimates (where they may be suppressed) as input
- Borrowing strength across groups could underestimate disparities
  - Outliers will be smoothed toward other larger groups or nearby time periods

Conclusions

- The eMKF macro can be used to produce model-based estimates of health outcomes for small subpopulations, to improve the availability of data for assessing disparities in small groups
- Large improvements in precision
  - Bigger gains for smaller subgroups
  - In some cases, equivalent to collecting 420% more data!
- Little increase in bias
  - Simulations show eMKF estimates generally closer to the true values than the subsample direct estimates, but not always
- Bayesian model average accurately captures trend form

Next Steps and Future Directions

- Further development to add new features
  - Jump in trend or discontinuity (e.g., survey design change, policy change)
  - Other types of trend specification (e.g., smoothing splines)
  - Other ideas?

- Any new features will be documented on the GitHub page
  - https://github.com/CDCgov/eMKF
    - Already includes examples from NHANES, NHIS, Vital Statistics, along with sample code to implement the macro
    - Can add more examples with different data systems

Contact us with questions or ideas:

- Lauren Rossen, lrossen@cdc.gov
- Makram Talih, veq0@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.