HHS Public Access

Author manuscript

J Surv Stat Methodol. Author manuscript; available in PMC 2025 January 08.

Published in final edited form as:

J Surv Stat Methodol. 2024 November ; 12(5): 1515–1530. doi:10.1093/jssam/smae036.

Real World Data Versus Probability Surveys for Estimating Health Conditions at the State Level

David A. Marker[*,1], **Charity Hilton**[2], **Jacob Zelko**[2], **Jon Duke**[2], **Deborah Rolka**[3], **Rachel Kaufmann**[3], **Richard Boyd**[2]

[1] Marker Consulting, Columbia, Maryland, United States of America

[2] Georgia Tech Research Institute, Atlanta, Georgia, United States of America

[3] National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America

Abstract

Government statistical offices worldwide are under pressure to produce statistics rapidly and for more detailed geographies, to compete with unofficial estimates available from web-based big data sources or from private companies. Commonly suggested sources of improved health information are electronic health records (EHRs) and medical claims data. These data sources are collectively known as real world data (RWD) because they are generated from routine health care processes, and they are available for millions of patients. It is clear that RWD can provide estimates that are more timely and less expensive to produce— but a key question is whether or not they are very accurate. To test this, we took advantage of a unique health data source that includes a full range of sociodemographic variables and compare estimates using all of those potential weighting variables, versus estimates derived when only age and sex are available for weighting (as is common with most RWD sources). We show that not accounting for other variables can produce misleading, and quite inaccurate, health estimates.

Keywords

nonprobability surveys; bias; data defect correlation; diabetes; electronic health records (EHRs); All of Us

1. Introduction

Government statistical offices around the world are under pressure to produce statistics rapidly and for more detailed geographies (CDC, 2024; Eurostat, 2019), to compete with unofficial estimates available from big data sources on the web or from private companies. For example, from Canada's chief statistician, "It's no longer enough to conduct surveys and look in the rear-view mirror at what happened. Citizens expect near real-time information..." (Arora, 2022). Commonly suggested sources of improved health information are electronic health records (EHRs) and medical claims data (Institute of Medicine, 2015). These data

* Corresponding author, MaryandDavidMarker@Gmail.com.

sources are collectively known as real world data (RWD) because they are generated from routine health care processes, and they are available for millions of patients. These sources may aggregate, for example, hospital visits, medications, diagnoses, test results, billings, and insurance data. The quality of RWD will vary across the world; in particular, countries such as the United States that do not have a centralized national health service are likely to have inconsistent coverage of the population in their RWD. As a result, these large datasets in the USA are not nationally representative, may be limited in terms of socioeconomic and demographic variables, and may lack the full health history needed to understand the context of the current test or diagnosis. For example, sources such as IQVIA, DartNet, and MarketScan may incorporate data from many states, but they typically only include providers located in a small number of communities within each state; Medicare data covers all of the United States, but only for those 65 and older. Thus, while it is clear that RWD can provide estimates that are more timely and less expensive to produce (although one should not underestimate the time and effort involved in combining data from multiple sources that were not collected under a common protocol), a key question is whether these estimates are sufficiently accurate.

Meng (2018) and Bradley et al. (2021) pointed out that if there is a correlation between who is found in the large data files and their characteristics, then the data defect correlation can cause large biases, making for very small effective sample sizes and very inaccurate estimates. This correlation is referred to as “selection bias” and “data not missing at random” in other literature. In the case of RWD, the probability of being included in the data is a combination of willingness to participate (by the person and/or their health system) and the likelihood of having certain tests conducted and then reported in the EHR.

Meng showed that the estimates’ accuracy is the product of three terms:

- Problem difficulty: the standard deviation of the variable being measured;
- Data quantity: $(1-f)/f = (N-n)/n$ (where f is the sampling fraction, n/N); and,
- Data quality: $\rho_{R,Y}$, the data defect correlation, which is the correlation between the indicator of inclusion, R, and the variable being measured, Y.

In probability surveys the last two terms cancel out, which is why statistics classes typically focus only on the first term, the problem difficulty. Big data sources are nonprobability samples of the population, so when measuring their accuracy, we have to include the other two terms as well.

The data quantity $(N-n)/n$ is not the same as the finite population correction $(N-n)/N$. The finite population correction is always less than or equal to 1, and it goes to zero as the sample size gets very large. But the data quantity can be much larger, although it too gets smaller as the sample size grows. As a simple example, if the data file contains one-tenth of the population (e.g., data on 26,000,000 adult Americans), the data quantity is equal to 9 (= $(N-0.1N)/0.1N = 0.9N/0.1N$).

The data defect correlation measures how unrepresentative the data are relative to the distribution of the variable being measured. In many situations this reflects whether

willingness to participate (e.g., nonresponse bias) is correlated with the outcome measure. But with EHRs the subject doesn't make a decision on whether to participate in the dataset, so we have to speak more generally about the likelihood of being included in the dataset (whether due to the patient interacting with the health care system, or the patient's health care network choosing to be included). Both Meng (2018) and Bradley et al. (2021) provide examples in which small values of this correlation (~0.006) can dramatically reduce the accuracy of estimates, in the latter case reducing a Facebook sample of 250,000 asking about vaccine uptake to an effective sample size of 10, even after weighting adjustments for age and sex. Yang et al. (2024) found similar results for the large Facebook sample in India and many other countries. They did, however, find that the same survey did much better at estimating changes in vaccine uptake across time, providing some hope that large non-probability surveys can be useful even when measurements of level are severely biased.

In that example, Bradley et al. show that by adjusting weighting for additional factors (such as race, education, political affiliation, and urban/rural location) one can dramatically reduce the correlation, and thus improve the accuracy from nonprobability sources.

In this paper we will examine the ability of such factors to improve the accuracy of electronic health records. While most EHRs have very limited factors to use in weighting adjustments, there is one such data source with additional variables that allows us to see how important such adjustments may be for health estimates. We examine whether such RWD can produce estimates for one important health indicator, diabetes prevalence, that are comparable to those produced by the probability survey currently relied upon by the U.S. Centers for Disease Control and Prevention (CDC).

2. Application to Electronic Health Records and Real World Data

2.1 Available Data

Health data on the U.S. population is collected through a series of probability surveys conducted by the CDC and other health agencies. These surveys produce high-quality health statistics, but they are expensive and take time to collect and process. Timeliness is a concern, particularly when information is urgently needed in the context of an emergency or emerging public health problem. Further, many of these surveys rest on respondents' self-reports and are therefore lacking detailed medical information and may have errors due to lack of accurate recollection. Finally, questionnaires are necessarily limited in length so as not to overburden respondents, which means that not all issues of interest can be included in each questionnaire. As a result, there is great interest to try and supplement the traditional surveys with routinely captured data such as EHRs and claims. Real world datasets are nonprobability collections of Americans, with certain groups likely to be overrepresented and others underrepresented. The hope is that with proper weighting (or calibration, propensity score modeling, or other techniques), biases from this non-representativeness can be minimized so that accurate estimates can be produced. Haneuse and Daniels (2016) point out that this hope hinges "on an overly simplistic view of the available/missing EHR data."

To evaluate the potential for such biases, we examined the ability of the All of Us database maintained by the National Institutes of Health (NIH) (https://allofus.nih.gov/) to provide improved state-level estimates of diabetes and diabetes control, compared to the Behavioral Risk Factor Surveillance System (BRFSS) conducted by CDC (https://www.cdc.gov/brfss/index.html). BRFSS is an annual telephone survey of 2,500 or more noninstitutionalized adults in each of the 50 states (and District of Columbia). In 2019, the total number of BRFSS respondents for the 50 states and DC was 409,760. Each state fields the same common "core" questionnaire as well as optional modules covering specific health topics of their choice. The sample selected for BRFSS supports unbiased direct estimation at the state level, but as with all surveys it suffers from significant nonresponse. It is also limited to asking respondents about their known conditions ("Has any medical provider told you that you have..."). In contrast, All of Us participants can contribute EHR data as well as complete a survey providing socioeconomic and demographic variables. Participants are a combination of volunteers across the country and targeted recruitment by NIH to include populations typically underrepresented in research. For these analyses we used BRFSS estimates for 2019 and All of Us enrollees from 2017–2022 for whom EHR data are available and include at least one office visit.

By capturing unreported cases and avoiding survey nonresponse, it is hoped that RWD might provide more precise estimates, but only if it avoids other forms of bias. While alternative RWD sources may have many times more participants, All of Us is the largest source with these additional variables. This provides a unique opportunity to identify some of the biases found across other RWD sources and examine how they impact state-level estimation.

All of Us is an ongoing effort of the NIH to collect EHR data from 1,000,000 Americans, who also respond to a socioeconomic questionnaire and provide blood samples for genomic analyses. While not representative of the U.S. population, All of Us will provide a unique combination of data sources on a large subset of the adult population. As of 2022, enrollment data were available for over 300,000 individuals, with approximately 280,000 also having EHR data available for analyses.

2.2 Sources of Potential Bias

Total Survey Error models (Groves et al., 2004) provide a convenient framework for comparing estimates from RWD such as All of Us with telephone surveys such as BRFSS. Table 1 highlights some of the error sources in both. As can be seen in the table, both data sources have multiple potential sources of bias. The largest sources for BRFSS are from nonresponse and the fact that many people with diabetes are unaware of their status. The largest sources for RWD are the unrepresentative nature of participants and the lack of covariates to use to adjust for this unrepresentativeness.

BRFSS's sampling frame includes all noninstitutionalized adults with a telephone, either landline or cell. Although each state is responsible for conducting its own surveys, it sometimes happens that the interviewer reaches a cell number that has moved to another state. In these instances, the core questionnaire is asked and the data are transferred to the number owner's new state of residence. This avoids any issue of bias arising from mobility

of respondents for the core. Diabetes diagnosis is included in the core questionnaire. However, optional modules are not asked of these respondents.

All of Us is open to all non-incarcerated adults in the United States who are able to give consent on their own. Because enrollment relies on individuals being aware of the database and choosing to volunteer for it, the participants are not necessarily representative of the general population.

2.3 Statistical Methods

We weighted the All of Us estimates using the full set of potential covariates and compared those estimates with BRFSS. Weights used age, sex, race/ethnicity, education, health insurance status, and income and were raked to control totals from the 2019 American Community Survey (ACS), conducted by the Census Bureau. (Due to difficulties with data collection during COVID-19, the 2020 ACS estimates are considered experimental, so we used the higher quality 2019 estimates.) Weighting is always a trade-off between bias reduction and increased variance due to differential weights. To minimize the variation in weights, we developed Python software that used different marginal controls for the weighting process (Georgia Tech Research Institute, 2023). The software generated weights using as many as six univariate marginals in the simplest case. It also generated weights with fewer univariate marginals and one or more two-dimensional marginals for the remaining variable pairs. We found that the model which reduced variability in weights the most included the interaction between sex and education. (For example, we controlled the number of females with a college degree, not just the number of females and the number of college graduates.) As a result, we had five marginal distributions used in the weighting[1]:

- Age: 18–29, 30–39, 40–49, 50–59, 60–69, and 70+;
- Race/ethnicity: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Hispanic, and Other;
- Health insurance status: yes or no/missing;
- Income: <$25,000, $25-$50,000, $50–100,000, and >$100,000; and,
- Sex-by-education: male and female crossed with: Less than high school, High school graduate, Some college, College graduate (or missing).

For the 17 states with more than 1,000 All of Us participants with health visit data, we use ACS totals for that state. For the remaining 34 states (treating DC as a state) with smaller numbers of participants, we aggregate the participants to each of the nine Census Divisions (https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf) and control their weights to the state ACS totals. For example, for Minnesota, we take all of the participants in the West North Central Division, treat them as if they were all from Minnesota, and weight them to the ACS control totals for Minnesota. We repeat this for the other six states in the West North Central Division as well. (This is the equivalent of using the synthetic estimator in small area estimation [NCHS, 1968], using proportions from the

[1]In some states it was necessary to further collapse some of these categories

division. It is possible to use a more complex model and improve upon synthetic estimates [Marker, 1999], but in this situation the improvement will be minimal (with only age and sex available, all models will be very simplistic) and will not affect the key findings.) In this way we are able to produce state-level estimates for all states, whose weighted totals match those in the ACS. (For specific states and divisions, we do some further collapsing of cells to keep a minimum cell size of at least 50 across all marginal distributions, which again limits the increase in variances.)

We then produce estimates of diabetes for each state to compare with those from BRFSS. We used the definition of diabetes from the SUPREME-DM study (Nichols et al., 2012):

- Diagnosis or some diabetes-specific medications; or,
- At least two instances of: A1c ≥ 6.5% or fasting plasma glucose ≥ 126 mg/dL.

This definition is somewhat more inclusive than BRFSS, which is limited to known diagnoses, so we expect somewhat larger estimates from All of Us. We also use All of Us to estimate the percentage of those with diabetes who have it under glycemic control. This is measured by examining the most recent A1c reading since the time of diagnosis. This metric cannot be estimated from survey questions, so it demonstrates the potential advantages of EHR data. We define control as follows:

- If the most recent A1c < 7% and since their initial diagnosis date—Under Control
- If the most recent A1c ≥ 7% and since their initial diagnosis date—Uncontrolled
- If they don’t have an A1c measurement since their initial diagnosis date—Indeterminate

While not the focus of this article, we are able to derive state-level estimates of control ranging from 32.5% in Utah to 61.5% for Massachusetts (with additional indeterminate cases in each state, some of which also are under control). The only survey-based official statistical estimate (from the National Health and Nutrition Examination Survey) is that nationally, 50.5% of people with diabetes have it under control (Fang et al., 2021). For the seven states in the West North Central Division, however, we could not produce estimates of control because none of the more than 100 participants with diabetes in those states had A1c measurements since their diagnosis.

To mimic the use of RWD sources lacking socioeconomic data, we reweighted All of Us restricting the weighting variables to only age and sex. While claims data may have race recorded, this variable is often missing or inconsistently recorded. Moreover, those who are reporting race are often quite unrepresentative, and there are no additional variables available for imputing missing race. By comparing these estimates to those from All of Us using the full set of weighting variables, we can examine the potential for bias in larger but less richly populated RWD sources.

3. Results and Interpretations

3.1 All of Us

As described above, we expect our All of Us estimates of diabetes prevalence to be somewhat higher than BRFSS since the SUPREME-DM definition goes beyond the BRFSS definition. We also expect that the state-level prevalence estimates will track relatively closely—although there will be differences since the proportion of people with a positive test for diabetes who have been told this by a health care professional varies from state to state (Danaei et al., 2009; Dwyer-Lindgren et al., 2016). If there are large discrepancies for particular states, it may imply that either All of Us participants are so unrepresentative, or data quality from that state is so poor, that even weighting for the six control variables cannot correct the biases.

Figure 1 shows the weighted 51 state estimates for BRFSS, ordered from lowest prevalence to highest, and the corresponding weighted All of Us estimates. In general, we do find a similar pattern across the states, with All of Us estimates somewhat higher (typically 3–5%) than those from BRFSS. There are, however, two particular exceptions.

There is one outlier in the bottom right (Figure 1), the estimate for the state of Louisiana. We have restricted the All of Us data to those participants for whom EHR data are available for at least one medical visit. Nationally, 60% of such participants have been diagnosed with at least one condition, but in Louisiana this is only 14%. Every other state has at least 36% diagnosed. This suggests (1) that the EHR data provided for Louisiana participants are not as complete compared to other states, and (2) that reliance on All of Us will produce an underestimate for any condition, even though there are more than 1,000 All of Us participants with EHR data in that state.

The second set of anomalous data on diabetes prevalence is the set of five state estimates in the top left of Figure 1, where more than 25% of the population would be estimated to have diabetes based on All of Us. These estimates are for Maine, New Hampshire, Vermont, Massachusetts, and Rhode Island. In the New England Census Division only Massachusetts and Connecticut had over 1,000 participants and thus produced direct estimates of diabetes prevalence. The other four state estimates were based on the entire set of participants in New England. It happens that 90% of those participants are from Massachusetts, so those four other states produce similar estimates to Massachusetts (how much they vary depends upon the differences in demographic and socioeconomic conditions from one state to another). Massachusetts' All of Us participants included a large percentage with diabetes, even after adjusting for these variables, though BRFSS indicates that state prevalence is far lower. This explains why this grouping of states have very similar, yet biased, estimates.

The impact on estimates of each variable used in weighting is presented in Figures 2 and 3. Figure 2 shows the marginal distributions for each of the six variables. As expected, age has the strongest predictive power, with diabetes prevalence increasing from 3.3% of 18–29-year-olds to 33.6% of those aged 70 or older. We also see that insurance coverage, race, and education are important predictors. Income has a general pattern, with higher income associated with lower prevalence of diabetes. But income is not as strong a predictor

as education. There is very little difference with the sex variable, but remember that we used the two-way distribution of sex-by-education for weighting. Figure 3 shows that the lowering of diabetes prevalence associated with increasing education is stronger for females than males.

These results in Figure 1 reveal that the All of Us dataset is generally able to produce state-level estimates of diabetes that follow similar patterns to BRFSS, but are typically 3–5% higher[5] as they incorporate some of the people who have diabetes but are not aware of it. (Undiagnosed cases, in which patients have not been to a medical provider for relevant tests, are still missing from All of Us.) For example, only four of the 20 states with the smallest BRFSS estimates have an All of Us estimate above 15%, while a majority of the remaining 31 states have All of Us estimates above that level. However, if either the quality of the EHR data in select states is incomplete, or those who participate are nonrepresentative, then relying on EHR data will provide misleading estimates, as were found in Louisiana and New England states.[6]

3.2 Applicability to Other Real World Data Sources

There are many other sources of real-world data, often with far more participants than All of Us. Most notable would be medical claims datasets, which may include tens of millions of patients. These sources typically lack participants' socioeconomic data such as income and education. Race/ethnicity data may be sparsely or inconsistently populated. Geographic data are typically present, either at the state level, metropolitan statistical area level, or 3-digit ZIP code level.

The recent work by Meng (2018) and Bradley et al. (2021) (discussed above) focused on the impact of not adjusting for bias in big data. In the latter's example of estimation of COVID-19 vaccination rates from early 2021, they demonstrated that when the likelihood of being in the dataset is correlated with the response (vaccination status), the accuracy of the estimates can be severely impacted if the variables to use in weighting to adjust for this correlation are not available. In other literature this is referred to as "selection bias", or "missing not at random" (MNAR). So, what might happen if we tried to estimate diabetes prevalence from larger RWD sources that reliably have only age and sex available for weighting?[7] To address this question, we re-weighted the All of Us data using only these two variables and examined differences in the estimates.

Figure 4 shows the state-level estimates produced from BRFSS and All of Us, when the latter is restricted to weighting by age and sex. We included the two-way interaction between these two in the weights. In general, we have the same patterns as found in Figure 1: the All of Us estimates are again generally higher than those from BRFSS, and again we have the problems of Louisiana and the five New England states. However, this time the estimates are different. In 31 of the 51 states the estimates changed by more than 1%; in Connecticut the estimated prevalence of diabetes increased by 9.8%, in South Carolina by 5.2%. (Change

[5]2019 BRFSS state-level standard errors range from 0.3% to 0.8% (CDC, 2022).
[6]We can hope that as the size of All of Us more than triples, reaching the 1,000,000 participant goal, both the quality of the EHR data and its representativeness will improve. But this will need to be evaluated before relying on it over a representative survey.
[7]From conversations the authors have had with representatives of other RWD and evaluations of their data.

relative to the full-model estimates are also large, 19 of the 51 states had their prevalence rates change by more than 10% from the full model.) Comparison with Figure 1 also reveals less of the overall pattern of higher estimates from All of Us states that have larger BRFSS estimates. While the true prevalence rates are unknown, given how correlated the covariates are with health outcome, we can assume that the full-model estimates that adjust for these covariates are in general more accurate than those using only age and sex.

When we look at the marginals for age-by-sex in Figure 5, we again see that age is a strong predictor, while sex is not. Female diabetes prevalence is higher for each age group through age 69, but they are lower than males' for the group 70 years and older.

But most interesting is comparing the estimates by sex in Figure 5 with those of the more complex model in Figure 2. When weighting for age, insurance coverage, race, income, and sex-by-education we estimate the prevalence of diabetes as:

Females 16.0% Males 16.7%

When we weight only on age-by-sex we have:

Females 17.1% Males 15.0%

With the simpler model we would report that females are more likely to have diabetes, but when we reduce sources of bias by adjusting for the nonrandom distributions of age, insurance, race, income, and education, we estimate that females are less likely to have diabetes. Moreover, in both datasets, the sample sizes are so large that both differences in sex are statistically significant in a two-sample t test. This demonstrates the limitations of RWD sources with limited available covariates. While their large sizes may enable generation of apparently very precise estimates, such results may also suffer from significant biases for which they cannot adjust, leading to spurious conclusions.

4. Summary

Our findings should serve as an important caveat regarding use of RWD that do not provide sufficient covariates to reduce the size of the data defect correlation. Big data sources can provide lots of cases, but the biases may swamp the reduction in sampling error. If the data do not come from a probability sample, even a small correlation between the response propensity and the variable of interest can dramatically reduce the accuracy of estimates.

It is vital to minimize such correlation through weighting, propensity score modeling, or other methods. But if strong covariates aren't provided with the dataset, none of these methods will provide the necessary adjustments.

Beesley and Mukherjee (2020) use simulations to demonstrate that raking and post-stratification weights will not necessarily eliminate selection bias if the likelihood of inclusion is related to the probability of having the disease. Our work demonstrates with a large EHR how the lack of available socio-demographic variables does indeed result in biased estimates, making estimates unreliable unless better covariates are available.

State-level estimates can be produced from EHRs and other RWD sources; in fact, one may be able to produce estimates for a wide range of health conditions and metrics that cannot readily be measured with self-reports. This potential is exciting as it would greatly increase our ability to monitor the health status of the population and do so relatively rapidly compared to population-based health examination surveys. However, the quality and completeness of a given RWD source may vary from state to state, so some estimates may be of much poorer quality than others. It is important to review the set of estimates for outliers (as shown in the example above where both Louisiana and the New England states' estimates are not at all consistent with other reported estimates) before automatically assuming that all estimates should be reported.

Hopefully, these findings can provide guidance to RWD consumers and producers as they seek to ensure that such data are fit for use in official government statistical and research purposes. In particular, inclusion of race/ethnicity data (as urged by IOM, 2015) would ensure a greater likelihood of RWD's utility in these settings, although the level of specificity may depend upon the size of the dataset and the racial composition of the population (for example, Asian-Americans are much more common in Washington state, Black Americans less common in the upper Midwest).

References

Arora A (2022, July). Modernizing government statistics for the 21st century. Amstat News. https://magazine.amstat.org/blog/2022/07/01/modernizing-government-statistics/

Beesley, and Mukherjee B (2020). Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. Biometrics. Doi: 10.1111/biom.13400

Bradley VC, Kuriwaki S, Isakov M, Sejdinovic D, Meng X-L, & Flaxman S (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. Nature, 600, 695–700. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8653636/ [PubMed: 34880504]

Centers for Disease Control and Prevention. 2024. CDC Data Modernization Efforts Accelerate Nation's Ability to Detect and Rapidly Respond to Health Threats. https://www.cdc.gov/media/releases/2024/p0411-CDC-data-modernization.html#:~:text=CDC%20Newsroom%20Releases-,CDC%20Data%20Modernization%20Efforts%20Accelerate%20Nation's%20Ability%20to,Rapidly%20Respond%20to%20Health%20Threats&text=Today%20the%20Centers%20for%20Disease,a%20companion%202023%20Lookback%20Report.

Centers for Disease Control and Prevention. 2022. Diabetes Data and Statistics. https://www.cdc.gov/diabetes/data/index.html

Danaei G, Friedman AB, Oza S, Murray CJL, & Ezzati M (2009). Diabetes prevalence and diagnosis in US states: analysis of health surveys. Population Health Metrics, 7, 16. 10.1186/1478-7954-7-16 [PubMed: 19781056]

Dwyer-Lindgren L, Mackenbach JP, van Lenthe FJ, Flaxman AD, Mokdad AH (2016). Diagnosed and undiagnosed diabetes prevalence by county in the U.S., 1999–2012. Diabetes Care, 39(9). 10.2337/dc16-0678

Eurostat (2019). Regulation (EU) 2019/1700 of the European Parliament and of the Council. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32019R1700

Fang M, Wang D, Coresh J, & Selvin E (2021). Trends in diabetes treatment and control in U.S. adults, 1999–2018. New England Journal of Medicine, 384, 2219–2228. doi: 10.1056/NEJMsa2032271 [PubMed: 34107181]

Georgia Tech Research Institute. (2023). MARKER software for iterative proportional fitting. https://apps.hdap.gatech.edu/t10docs/

Groves RM, Fowler FJ Jr., Couper MP, Lepkowski JM, Singer E, and Tourangeau R (2004). Survey Methodology. John Wiley & Sons.

Haneuse S and Daniels M (2016). A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? EGEMS (Washington DC). 4(1):1203. doi: 10.13063/2327-9214.1203 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013936/ [PubMed: 27668265]

Institute of Medicine (2015). Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records. Washington (DC): National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK269330/

Marker DA (1999). Organization of small area estimators using a generalized linear regression framework. Journal of Official Statistics, 15(1), 1–24. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/organization-of-small-area-estimators-using-a-generalized-linear-regression-framework..pdf

Meng X-L (2018, June). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics, 12(2), 685–726. doi: 10.1214/18-AOAS1161SF

National Center for Health Statistics (1968). Synthetic state estimates of disability (P.H.S. Publication No. 1759). Government Printing Office.

Nichols GA, Desai J, Lafata JE, et al. (2012). Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: The SUPREME-DM project. Preventing Chronic Disease, 9, E110. doi: 10.5888/pcd9.110311 [PubMed: 22677160]

Yang Y, Dempsey W, Han P, Deshmukh Y, Richardson S, Tom B, and Mukherjee B (2024). Exploring the big data paradox for various estimands using vaccination data from the global COVID-19 Trends and Impact Survey (CTIS). Science Advances 10:22. https://www.science.org/doi/10.1126/sciadv.adj0266

Statement of Significance

Government statistical offices worldwide are under pressure to produce statistics rapidly and for more detailed geographies, to compete with unofficial estimates available from web-based big data sources or from private companies. Commonly suggested sources of improved health information are electronic health records (EHRs) and medical claims data. Such real-world data (RWD) are available for millions of patients. It is clear that RWD can provide estimates that are more timely and less expensive to produce— a key question is whether or not they are very accurate.

We examine the ability of the All of Us RWD maintained by the National Institutes of Health to improve state-level estimates of diabetes compared to those currently produced by the Centers for Disease Control and Prevention (CDC) from representative high-quality surveys. Unlike most RWD sources, All of Us includes demographic and socioeconomic data on participants, allowing for minimizing biases in the unrepresentative RWD. We then restrict All of Us to the limited set of such data common in other RWD, and show how different the estimates are, indicating that most other RWD sources produce biased estimates that cannot be relied upon to improve upon government survey data.

We are not aware of any previous use of All of Us (or any alternative source) to explicitly evaluate bias in RWD. This unique use of All of Us demonstrates the biases that can exist in RWD and cannot be adjusted for due to the limited set of available covariates.

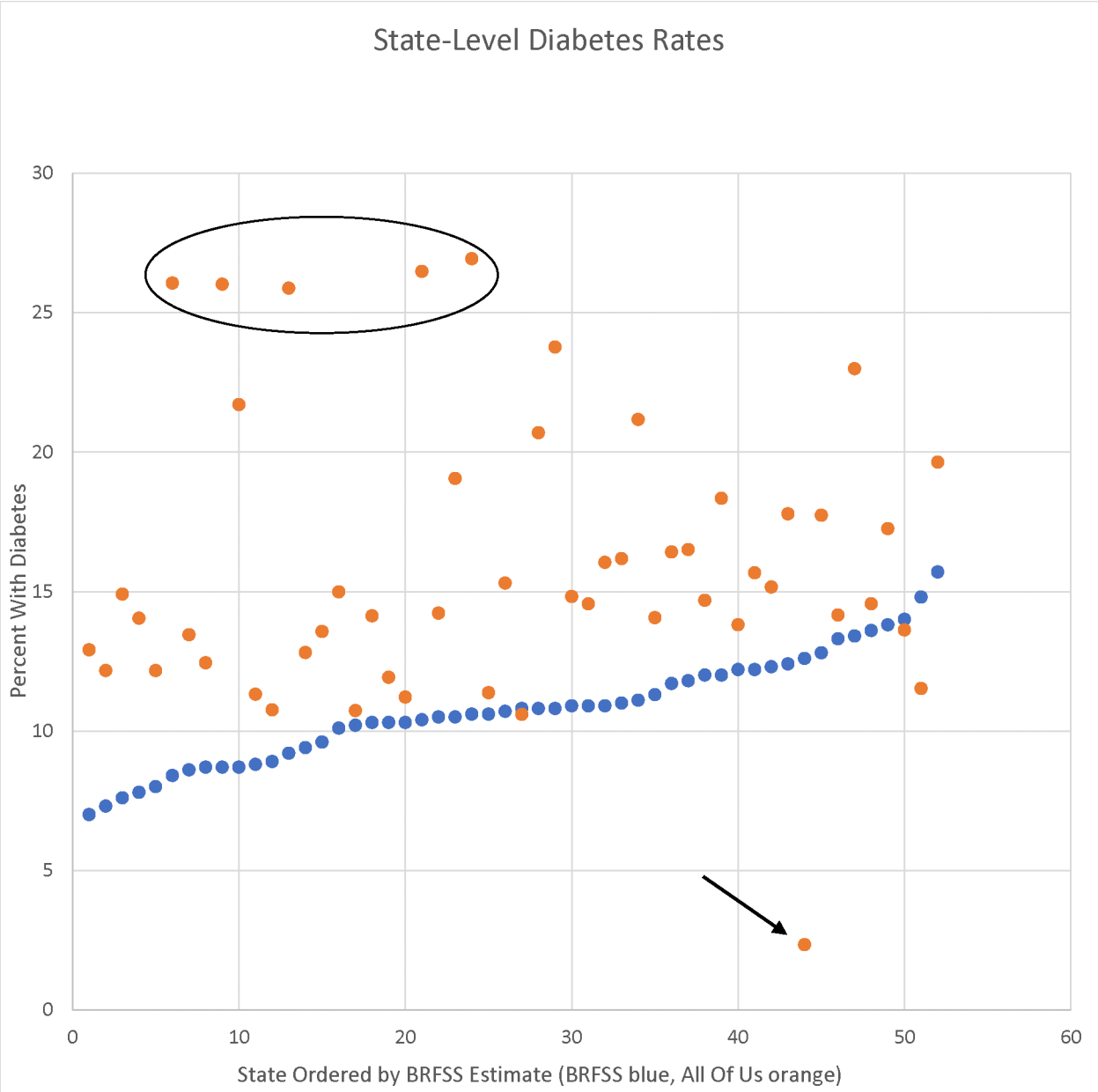

Figure 1.
State-Level Weighted Estimates of Diabetes from BRFSS and All of Us (ellipse covers Northeast Division and arrow points to Louisiana)[2]

[2]Data are for all All of Us enrollees from 2017–2022 for whom EHR data are available and include at least one office visit.

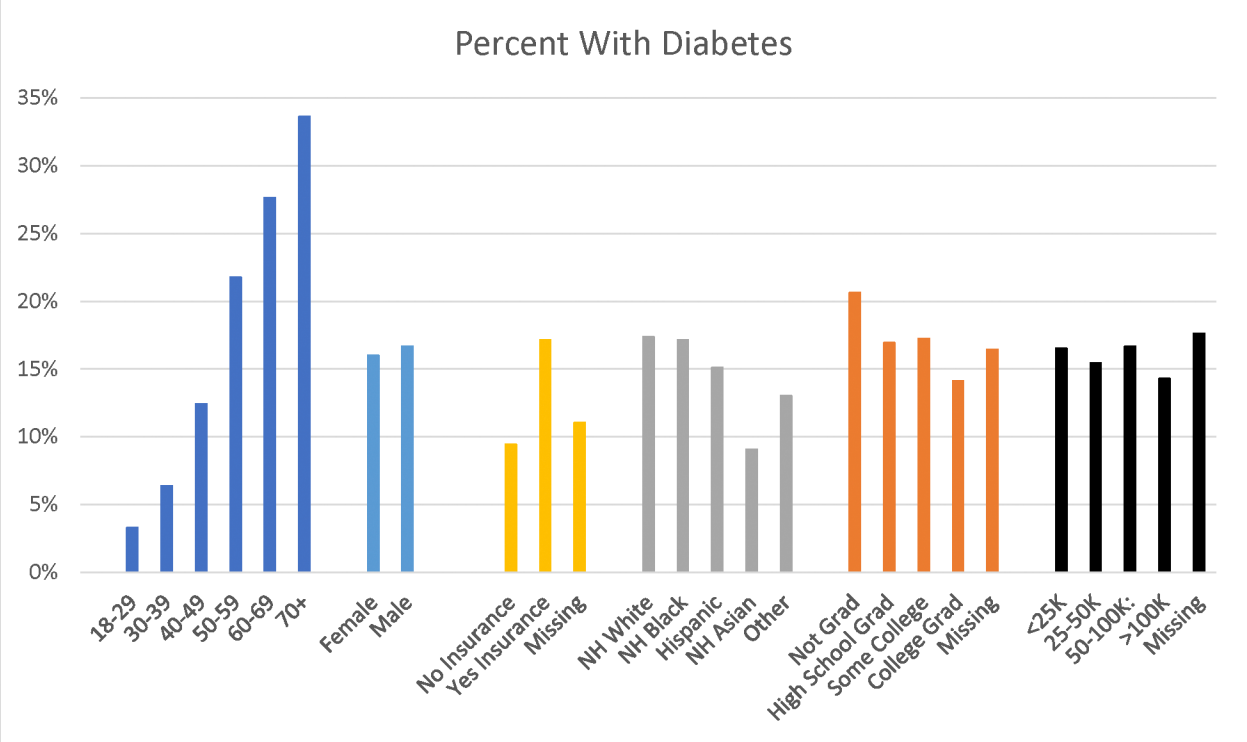

Figure 2.
Weighted Estimated Percent with Diabetes for Each Marginal Distribution: All of Us[3]

[3]Data are for all All of Us enrollees from 2017–2022 for whom EHR data are available and include at least one office visit.

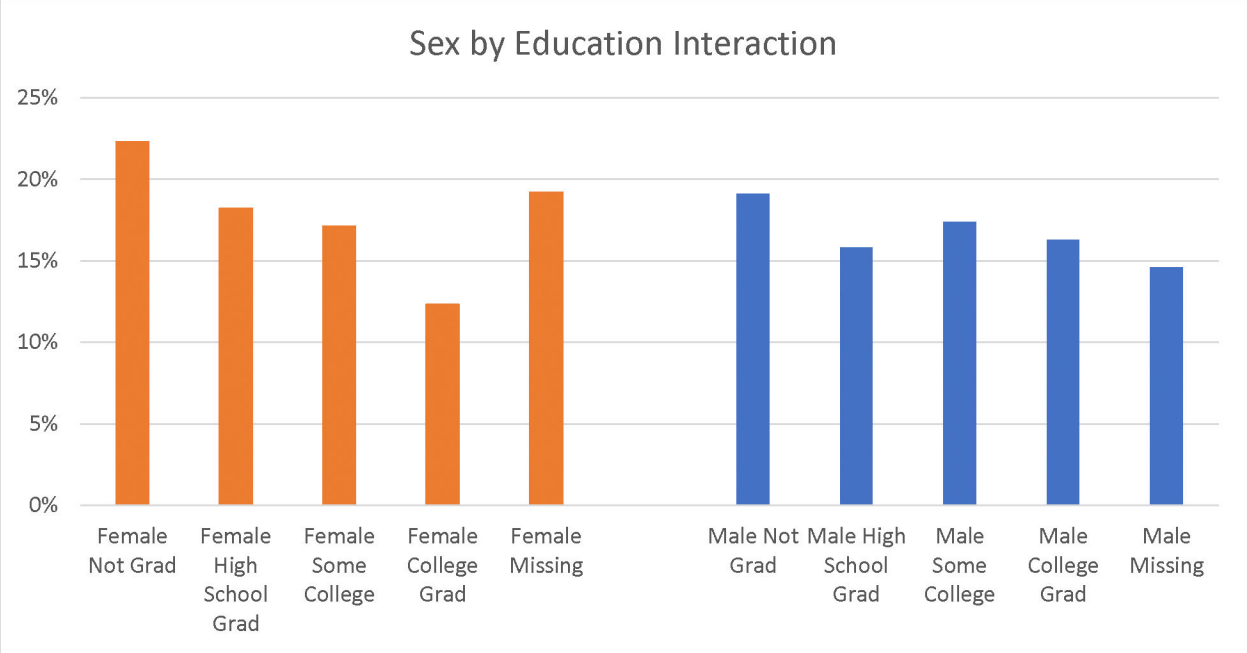

Figure 3.
Weighted Estimated Percent with Diabetes for Sex-by-Education Distribution: All of Us[4]

[4]Data are for all All of Us enrollees from 2017–2022 for whom EHR data are available and include at least one office visit.

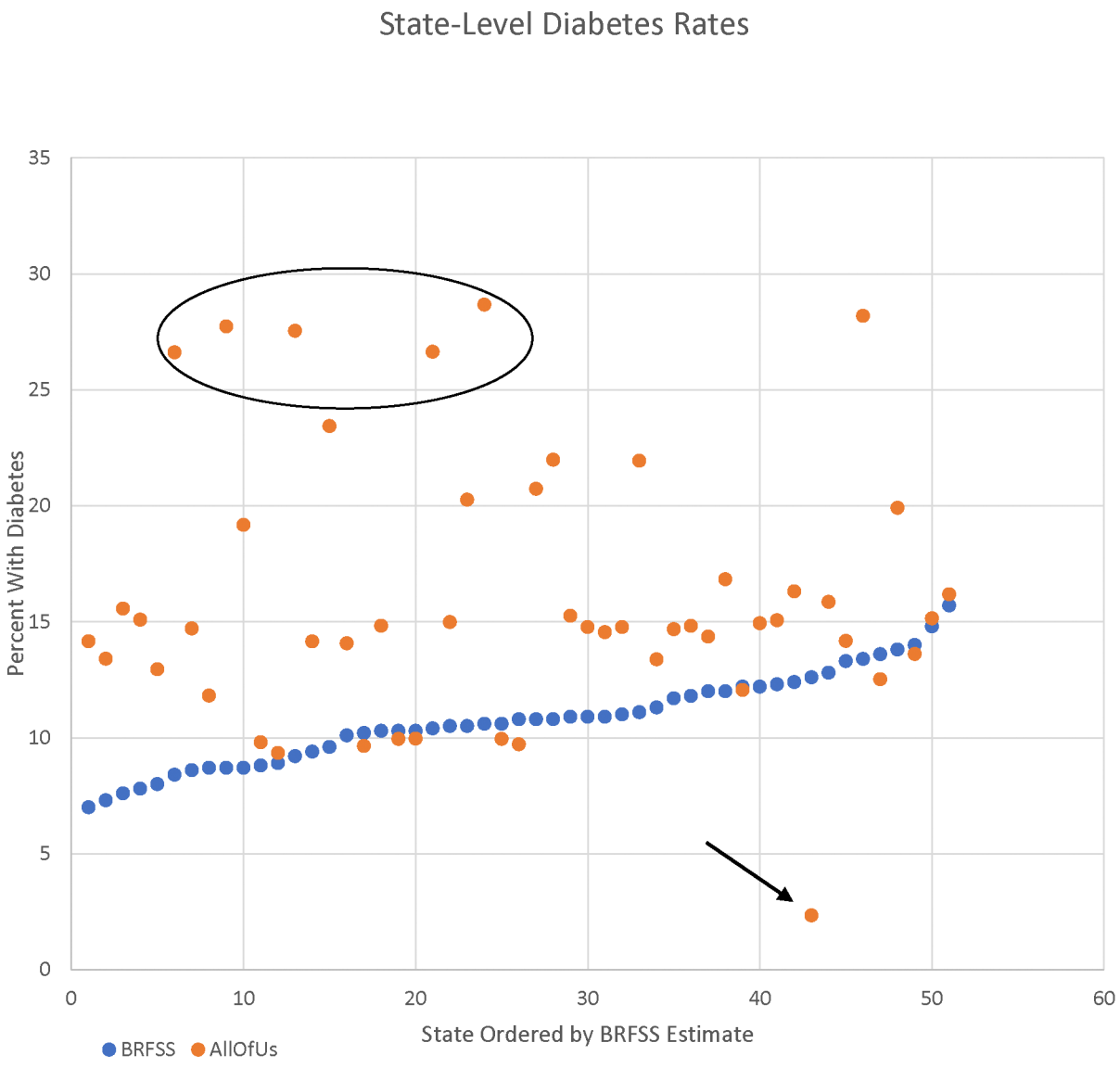

Figure 4.
State-Level Weighted Estimates of Diabetes from BRFSS and All of Us, Weighting on Age-by-Sex (ellipse covers Northeast Division and arrow points to Louisiana)[8]

[8]Data are for all All of Us enrollees from 2017–2022 for whom EHR data are available and include at least one office visit.

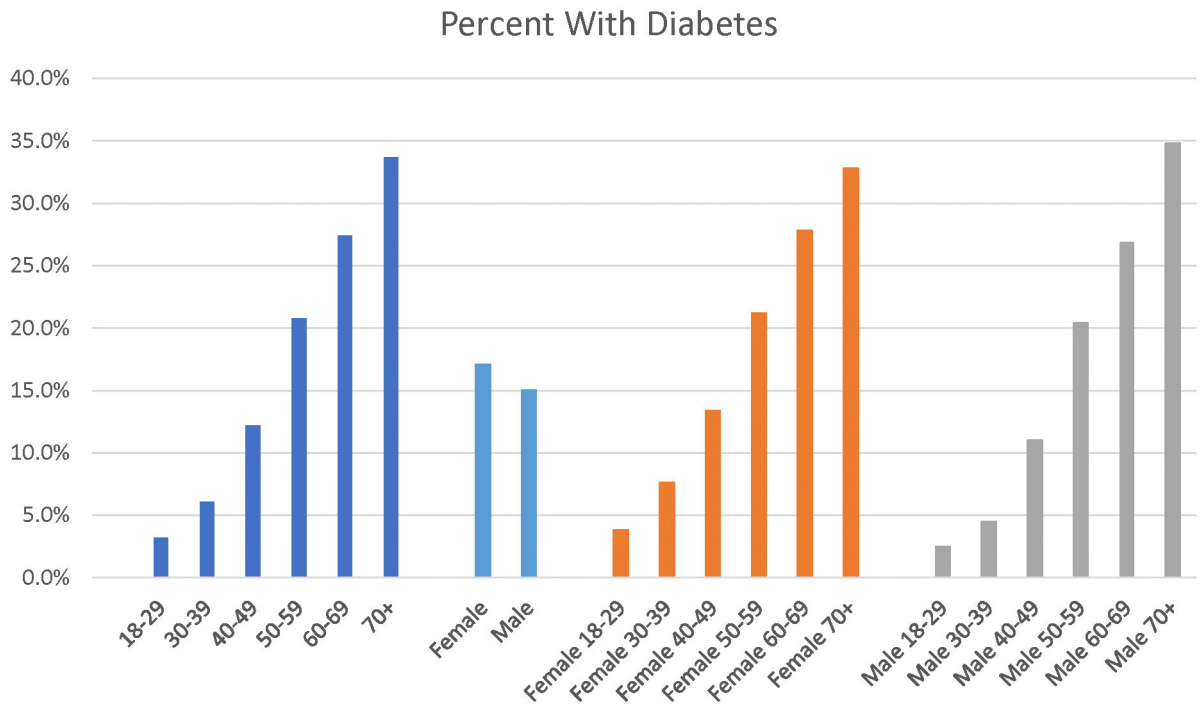

Figure 5.
Weighted Estimated Percent with Diabetes by Age and Sex When Weighting All of Us on Only These Variables[9]

[9]Data are for all All of Us enrollees from 2017–2022 for whom EHR data are available and include at least one office visit.

Table 1.

Sources of error in BRFSS and RWD in estimating prevalence of diabetes

	BRFSS	RWD
Capturing full range of diabetes	Only cases previously identified to respondent by a medical professional	Includes cases unknown to respondent, if a medical test has been conducted during the years included in the RWD
Coverage of Population	Only missing those without telephones (<2% of households in the USA)	Those who volunteer for All of Us. Only participants in health care plans included in other RWD. Not all regularly tested for diabetes
Unit nonresponse	Telephone surveys have high NR rates, limited covariates available for adjustment	Minimal
Item nonresponse	Minimal	Can be large and systematic depending on what information is recorded in the EHR
Ability for estimates to adjust for non-representativeness	Wide range of variables to adjust for item NR, but very limited ability to adjust for unit NR. No ability to adjust for unknown cases of diabetes	Very limited ability (except for All of Us) to adjust for any non-representativeness