Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

https://doi.org/10.1038/s41586-021-04198-4

Received: 18 June 2021

Accepted: 29 October 2021

Published online: 08 December 2021

Check for updates

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5]✉

Surveys are a crucial tool for understanding public opinion and behaviour, and their accuracy depends on maintaining statistical representativeness of their target populations by minimizing biases from all sources. Increasing data size shrinks confidence intervals but magnifies the effect of survey bias: an instance of the Big Data Paradox[1]. Here we demonstrate this paradox in estimates of first-dose COVID-19 vaccine uptake in US adults from 9 January to 19 May 2021 from two large surveys: Delphi–Facebook[2,3] (about 250,000 responses per week) and Census Household Pulse[4] (about 75,000 every two weeks). In May 2021, Delphi–Facebook overestimated uptake by 17 percentage points (14–20 percentage points with 5% benchmark imprecision) and Census Household Pulse by 14 (11–17 percentage points with 5% benchmark imprecision), compared to a retroactively updated benchmark the Centers for Disease Control and Prevention published on 26 May 2021. Moreover, their large sample sizes led to miniscule margins of error on the incorrect estimates. By contrast, an Axios–Ipsos online panel[5] with about 1,000 responses per week following survey research best practices[6] provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework[1] to explain the inaccuracy in the three surveys. We then analyse the implications for vaccine hesitancy and willingness. We show how a survey of 250,000 respondents can produce an estimate of the population mean that is no more accurate than an estimate from a simple random sample of size 10. Our central message is that data quality matters more than data quantity, and that compensating the former with the latter is a mathematically provable losing proposition.

Governments, businesses and researchers rely on survey data to inform the provision of government services[7], steer business strategy and guide the response to the COVID-19 pandemic[8,9]. With the ever-increasing volume and accessibility of online surveys and organically collected data, the line between traditional survey research and Big Data is becoming increasingly blurred[10]. Large datasets enable the analysis of fine-grained subgroups, which are in high demand for designing targeted policy interventions[11]. However, counter to common intuition[12], larger sample sizes alone do not ensure lower error. Instead, small biases are compounded as sample size increases[1].

We see initial evidence of this in the discrepancies in estimates of first-dose COVID-19 vaccine uptake, willingness and hesitancy from three online surveys in the US. Two of them—Delphi–Facebook's COVID-19 symptom tracker[2,3] (around 250,000 responses per week and with over 4.5 million responses from January to May 2021) and the Census Bureau's Household Pulse survey[4] (around 75,000 responses per survey wave and with over 600,000 responses from January to May 2021)—have large enough sample sizes to render standard uncertainty intervals negligible; however, they report significantly different estimates of vaccination behaviour with nearly identically worded questions (Table 1). For example, Delphi–Facebook's state-level estimates for willingness to receive a vaccine from the end of March 2021 are 8.5 percentage points lower on average than those from the Census Household Pulse (Extended Data Fig. 1a), with differences as large as 16 percentage points.

The US Centers for Disease Control and Prevention (CDC) compiles and reports vaccine uptake statistics from state and local offices[13]. These figures serve as a rare external benchmark, permitting us to compare survey estimates of vaccine uptake to those from the CDC. The CDC has noted the discrepancies between their own reported vaccine uptake and that of the Census Household Pulse[14,15], and we find even larger discrepancies between the CDC and Delphi–Facebook data (Fig. 1a). By contrast, the Axios–Ipsos Coronavirus Tracker[5] (around 1,000 responses per wave, and over 10,000 responses from January to May 2021) tracks the CDC benchmark well. None of these surveys use the CDC benchmark to adjust or assess their estimates of vaccine uptake, thus by examining patterns in these discrepancies, we can infer each survey's accuracy and statistical representativeness, a nuanced concept that is critical for the reliability of survey findings[16–19].

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Department of Political Science, Stanford University, Stanford, CA, USA. [3]Harvard College, Harvard University, Cambridge, MA, USA. [4]Department of Statistics, Harvard University, Cambridge, MA, USA. [5]Department of Computer Science, University of Oxford, Oxford, UK. [6]These authors contributed equally: Valerie C. Bradley, Shiro Kuriwaki. ✉e-mail: seth.flaxman@cs.ox.ac.uk

Table 1 | Comparison of survey designs

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Recruitment mode	Address-based mail sample to Ipsos KnowledgePanel	SMS and email	Facebook Newsfeed
Interview mode	Online	Online	Online
Average size	1,000/wave	75,000/wave	250,000/week
Sampling frame	Ipsos KnowledgePanel; internet/ tablets provided to ~5% of panelists who lack home internet	Census Bureau's Master Address File (individuals for whom email / phone contact information is available)	Facebook active users
Vaccine uptake question	"Do you personally know anyone who has already received the COVID-19 vaccine?"	"Have you received a COVID-19 vaccine?"	"Have you had a COVID-19 vaccination?"
Vaccine uptake definition	"Yes, I have received the vaccine"	"Yes"	"Yes"
Other vaccine uptake response options	"Yes, a member of my immediate family", "Yes, someone else", "No"	"No"	"No", "I don't know"
Weighting variables	Gender by age, race, education, Census region, metropolitan status, household income, partisanship.	Education by age by sex by state, race/ethnicity by age by sex by state, household size	Stage 1: age, gender "other attributes which we have found in the past to correlate with survey outcomes" to FAUB; Stage 2: state by age by gender

Comparison of key design choices across the Axios–Ipsos, Census Household Pulse and Delphi–Facebook studies. All surveys target the US adult population. See Extended Data Table 1 for additional comparisons and Methods for additional implementation details.

The Big Data Paradox in vaccine uptake

We focus on the Delphi–Facebook and Census Household Pulse surveys because their large sample sizes (each greater than 10,000 respondents[20]) present an opportunity to examine the Big Data Paradox[1] in surveys. The Census Household Pulse is an experimental product designed to rapidly measure pandemic-related behaviour. Delphi–Facebook has stated that the intent of their survey is to make comparisons over space, time and subgroups, and that point estimates should be interpreted with caution[3]. However, despite these intentions, Delphi–Facebook has reported point estimates of vaccine uptake in its own publications[11,21].

Delphi–Facebook and Census Household Pulse surveys persistently overestimate vaccine uptake relative to the CDC's benchmark (Fig. 1a) even taking into account Benchmark Imprecision (Fig. 1b) as explained in 'Decomposing Error in COVID Surveys'. Despite being the smallest survey by an order of magnitude, the estimates of Axios–Ipsos track well with the CDC rates (Fig. 1a), and their 95% confidence intervals contain the benchmark estimate from the CDC in 10 out of 11 surveys (an empirical coverage probability of 91%).

One might hope that estimates of changes in first-dose vaccine uptake are correct, even if each snapshot is biased. However, errors have increased over time, from just a few percentage points in January 2021 to Axios-Ipsos' 4.2 percentage points [1–7 percentage points with 5% benchmark imprecision (BI)], Census Household Pulse's 14 percentage points [5% BI: 11–17] and Delphi-Facebook's 17 percentage points [5% BI: 14–20] by mid-May 2021 (Fig. 1b). For context, for a state that is near the herd immunity threshold (70–80% based on recent estimates[22]), a discrepancy of 10 percentage points in vaccination rates could be the difference between containment and uncontrolled exponential growth in new SARS-CoV-2 infections.

Conventional statistical formulas for uncertainty further mislead when applied to biased big surveys because as sample size increases, bias (rather than variance) dominates estimator error. Figure 1a shows 95% confidence intervals for vaccine uptake based on the reported sampling standard errors and weighting design effects of each survey[23]. Axios–Ipsos has the widest confidence intervals, but also the smallest design effects (1.1–1.2), suggesting that its accuracy is driven more by minimizing bias in data collection rather than post-survey adjustment. The 95% confidence intervals of Census Household Pulse are widened by large design effects (4.4–4.8) but they are still too narrow to include the true rate of vaccine uptake in almost all survey waves. The confidence intervals for Delphi–Facebook are extremely small, driven by large sample size and moderate design effects (1.4–1.5), and give us a negligible chance of being close to the truth.

One benefit of such large surveys might be to compare estimates of spatial and demographic subgroups[24–26]. However, relative to the CDC's contemporaneously reported state-level estimates, which did not include retroactive corrections, Delphi–Facebook and Census Household Pulse

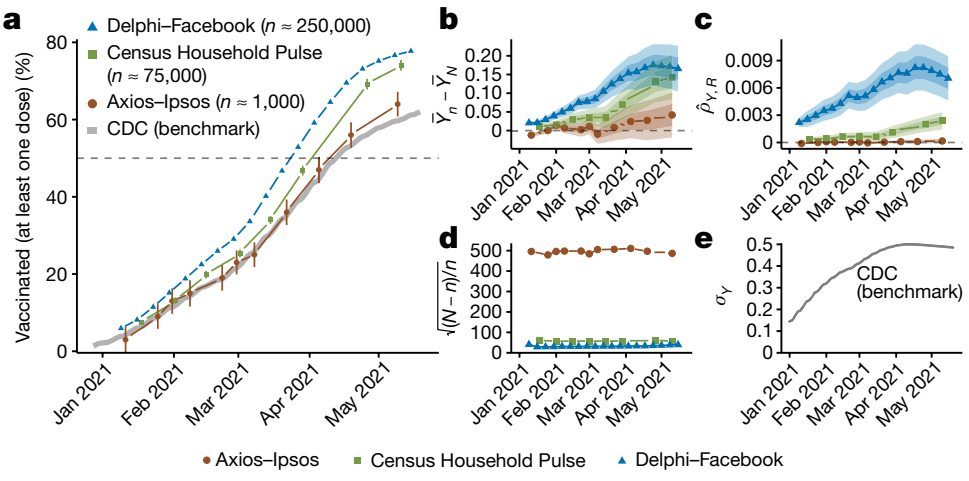


Fig 1 | Errors in estimates of vaccine uptake. **a**, Estimates of vaccine uptake for US adults in 2021 compared to CDC benchmark data, plotted by the end date of each survey wave. Points indicate each study's weighted estimate of first-dose vaccine uptake, and intervals are 95% confidence intervals using reported standard errors and design effects. Delphi–Facebook has $n = 4{,}525{,}633$ across 19 waves, Census Household Pulse has $n = 606{,}615$ across 8 waves and Axios–Ipsos has $n = 11{,}421$ across 11 waves. Delphi–Facebook's confidence intervals are too small to be visible. **b**, Total error $\bar{Y}_n - \bar{Y}_N$. **c**, Data defect correlation $\hat{\rho}_{Y,R}$. **d**, Data scarcity $\sqrt{(N-n)/n}$. **e**, Inherent problem difficulty σ_Y. Shaded bands represent scenarios of ±5% (darker) and ±10% (lighter) imprecision in the CDC benchmark relative to reported values (points). **b**–**e** comprise the decomposition in equation (1).

overestimated CDC state-level vaccine uptake by 16 and 9 percentage points, respectively (Extended Data Fig. 1g, h) in March 2021, and by equal or larger amounts by May 2021 (Extended Data Fig. 2g, h). Relative estimates were no better than absolute estimates in March of 2021: there is little agreement in a survey's estimated state-level rankings with the CDC (a Kendall rank correlation of 0.31 for Delphi–Facebook in Extended Data Fig. 1i and 0.26 for Census Household Pulse in Extended Data Fig. 1j) but they improved in May of 2021 (correlations of 0.78 and 0.74, respectively, in Extended Data Fig. 2i, j). Among 18–64-year-olds, both Delphi–Facebook and Census Household Pulse overestimate uptake, with errors increasing over time (Extended Data Fig. 6).

These examples illustrate a mathematical fact. That is, when biased samples are large, they are doubly misleading: they produce confidence intervals with incorrect centres and substantially underestimated widths. This is thev Big Data Paradox: "the bigger the data, the surer we fool ourselves"[1] when we fail to account for bias in data collection.

A framework for quantifying data quality

Although it is well-understood that traditional confidence intervals capture only survey sampling errors[27] (and not total error), the traditional survey framework lacks analytic tools for quantifying nonsampling errors separately from sampling errors. A previously formulated statistical framework[1] permits us to exactly decompose the total error of a survey estimate into three components:

$$\text{Total error} = \text{Data quality defect} \times \text{Data scarcity} \times \text{Inherent problem difficulty} \quad (1)$$

This framework has been applied to COVID-19 case counts[28] and election forecasting[29]. Its full application requires ground-truth benchmarks or their estimates from independent sources[1].

Specifically, the 'total error' is the difference between the observed sample mean $\overline{Y}_n$ as an estimator of the ground truth, the population mean $\overline{Y}_N$. The 'data quality defect' is measured using $\hat{\rho}_{Y,R}$, called the 'data defect correlation' (ddc)[1], which quantifies total bias (from any source), measured by the correlation between the event that an individual's response is recorded and its value, Y. The effect of data quantity is captured by 'data scarcity', which is a function of the sample size n and the population size N, measured as $\sqrt{(N-n)/n}$, and hence what matters for error is the relative sample size—that is, how close n is to N—rather than the absolute sample size n. The third factor is the 'inherent problem difficulty', which measures the population heterogeneity (via the standard deviation σ_Y of Y), because the more heterogeneous a population is, the harder it is to estimate its average well. Mathematically, equation (1) is given by $\overline{Y}_n - \overline{Y}_N = \hat{\rho}_{Y,R} \times \sqrt{(N-n)/n} \times \sigma_Y$. This expression was inspired by the Hartley–Ross inequality for biases in ratio estimators[30]. More details on the decomposition are provided in 'Calculation and interpretation of ddc' in the Methods, in which we also present a generalization for weighted estimators.

Decomposing error in COVID surveys

Although the ddc is not directly observed, COVID-19 surveys present a rare case in which it can be deduced because all of the other terms in equation (1) are known (see 'Calculation and interpretation of ddc' in the Methods for an in-depth explanation). We apply this framework to the aggregate error shown in Fig. 1b, and the resulting components of error from the right-hand side of equation (1) are shown in Fig. 1c–e.

We use the CDC's report of the cumulative count of first doses administered to US adults as the benchmark[8,13], $\overline{Y}_N$. This benchmark time series may be affected by administrative delays and slippage in how the CDC centralizes information from states[31–34]. The CDC continuously updates their entire time series retroactively for such delays as they are reported. But to account for potentially unreported delays, we present our results with Benchmark Imprecision (BI) in case the CDC's numbers from our study period, 9 January to 26 May 2021, as reported on 26 May by the CDC suffer from ±5% and ±10% imprecision. These scenarios were chosen on the basis of analysis of the magnitude by which the CDC's initial estimate for vaccine uptake by a particular day increases as the CDC receives delayed reports of vaccinations that occurred on that day (Extended Data Fig. 3, Supplementary Information A.2). That said, these scenarios may not capture latent systemic issues that affect CDC vaccination reporting.

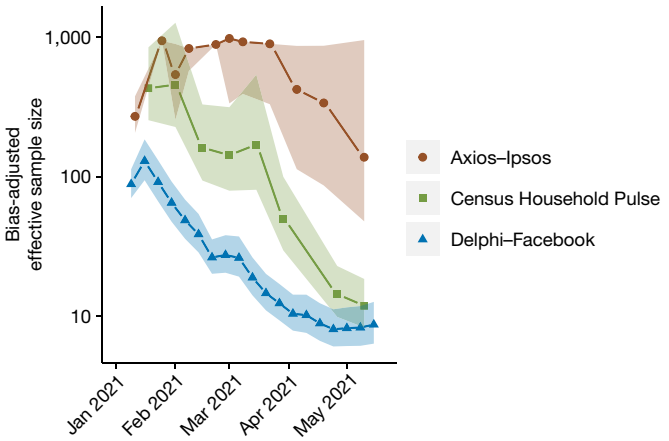


Fig 2 | Bias-adjusted effective sample size. An estimate's bias-adjusted effective sample size (different from the classic Kish effective sample size) is the size of a simple random sample that would have the same MSE as the observed estimate. Effective sample sizes are shown here on the $\log_{10}$ scale. The original sample size was $n = 4{,}525{,}633$ across 19 waves for Delphi–Facebook, $n = 606{,}615$ across 8 waves for Census Household Pulse and $n = 11{,}421$ across 11 waves for Axios–Ipsos. Shaded bands represent scenarios of ±5% benchmark imprecision in the CDC benchmark.

The total error of each survey's estimate of vaccine uptake (Fig. 1b) increases over time for all studies, most markedly for Delphi–Facebook. The data quality defect, measured by the ddc, also increases over time for Census Household Pulse and for Delphi–Facebook (Fig. 1c). The ddc for Axios–Ipsos is much smaller and steady over time, consistent with what one would expect from a representative sample. The data scarcity, $\sqrt{(N-n)/n}$, for each survey is roughly constant across time (Fig. 1d). Inherent problem difficulty is a population quantity common to all three surveys that peaks when the benchmark vaccination rate approaches 50% in April 2021 (Fig. 1e). Therefore, the decomposition suggests that the increasing error in estimates of vaccine uptake in Delphi–Facebook and Census Household Pulse is primarily driven by increasing ddc, which captures the overall effect of the bias in coverage, selection and response.

Equation (1) also yields a formula for the bias-adjusted effective sample size n_{eff}, which is the size of a simple random sample that we would expect to exhibit the same level of mean squared error (MSE) as what was actually observed in a given study with a given ddc. Unlike the classical effective sample size[23], this quantity captures the effect of bias as well as that of an increase in variance from weighting and sampling. For details of this calculation, see 'Error decomposition with survey weights' in the Methods.

For estimating the US vaccination rate, Delphi–Facebook has a bias-adjusted effective sample size of less than 10 in April 2021, a 99.99% reduction from the raw average weekly sample size of 250,000 (Fig. 2). The Census Household Pulse is also affected by over 99% reductions in effective sample size by May 2021. A simple random sample would have controlled estimation errors by controlling ddc. However, once this control is lost, small increases in ddc beyond what is expected in simple random samples can result in marked reductions of effective sample sizes for large populations[1].

Table 2 | Composition of survey respondents by educational attainment and race/ethnicity

	Composition of US adults							Survey estimates		
	Axios–Ipsos		**Household Pulse**		**Delphi–Facebook**		**ACS**	**Household Pulse**		
	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
Education										
High school	35%	39%	14%	39%	19%	21%	39%	39%	40%	21%
Some college	29	30	32	30	36	36	30	44	38	18
Four-year college	19	17	29	17	25	25	19	54	36	10
Post-graduate	17	14	26	13	20	18	11	67	26	7
Race/ethnicity										
White	71%	63%	75%	62%	74%	68%	60%	50%	33%	17%
Black	10	12	7	11	6	6	12	42	39	19
Hispanic	11	16	10	17	11	16	16	38	48	14
Asian			5	5	2	3	6	51	43	5

Axios–Ipsos: wave ending 22 March 2021, n = 995. Census Household Pulse: wave ending 29 March 2021, n = 76,068. Delphi–Facebook: wave ending 27 March 2021, n = 181,949. Benchmark uses the 2019 US Census American Community Survey (ACS), composed of roughly 3 million responses. The rightmost column shows estimates of vaccine uptake (Vax), willingness (Will) and hesitancy (Hes) from the Census Household Pulse of the same wave.

Comparing study designs

Understanding why bias occurs in some surveys but not others requires an understanding of the sampling strategy, modes, questionnaire and weighting scheme of each survey. Table 1 compares the design of each survey (more details in 'Additional survey methodology' in the Methods, Extended Data Table 1).

All three surveys are conducted online and target the US adult population, but vary in the methods that they use to recruit respondents[35]. The Delphi–Facebook survey recruits respondents from active Facebook users (the Facebook active user base, or FAUB) using daily unequal-probability stratified random sampling[2]. The Census Bureau uses a systematic random sample to select households from the subset of the master address file (MAF) of the Census for which they have obtained either cell phone or email contact information (approximately 81% of all households in the MAF)[4].

In comparison, Axios–Ipsos relies on inverse response propensity sampling from Ipsos' online KnowledgePanel. Ipsos recruits panellists using an address-based probabilistic sample from USPS's delivery sequence file (DSF)[5]. The DSF is similar to the MAF of the Census. Unlike the Census Household Pulse, potential respondents are not limited to the subset for whom email and phone contact information is available. Furthermore, Ipsos provides internet access and tablets to recruited panellists who lack home internet access. In 2021, this 'offline' group typically comprises 1% of the final survey (Extended Data Table 1).

All three surveys weight on age and gender; that is, assign larger weights to respondents of underrepresented age by gender subgroups and smaller weights to those of overrepresented subgroups[2,4,5] (Table 1). Axios–Ipsos and Census Household Pulse also weight on education and race and/or ethnicity (hereafter, race/ethnicity). Axios–Ipsos additionally weights to the composition of political partisanship measured by "recent ABC News/Washington Post telephone polls"[5] in 6 of the 11 waves we study. Education—a known correlate of propensity to respond to surveys[36] and social media use[37] are notably absent from Delphi–Facebook's weighting scheme, as is race/ethnicity. As noted before, none of the surveys use the CDC benchmark to adjust or assess estimates of vaccine uptake.

Explanations for error

Table 2 illustrates some consequences of these design choices. Axios–Ipsos samples mimic the actual breakdown of education attainment among US adults even before weighting, whereas those of Census Household Pulse and Delphi–Facebook do not. After weighting, Axios–Ipsos and Census Household Pulse match the population benchmark, by design. Delphi–Facebook does not explicitly weight on education, and hence the education bias persists in their weighted estimates: those without a college degree are underrepresented by nearly 20 percentage points. The case is similar for race/ethnicity. Delphi–Facebook's weighting scheme does not adjust for race/ethnicity, and hence their weighted sample still overrepresents white adults by 8 percentage points, and underrepresents the proportions of Black and Asian individuals by around 50% of their size in the population (Table 2).

The overrepresentation of white adults and people with college degrees explains part of the error of Delphi–Facebook. The racial groups that Delphi–Facebook underrepresents tend to be more willing and less vaccinated in the samples (Table 2). In other words, reweighting the Delphi–Facebook survey to upweight racial minorities will bring willingness estimates closer to Household Pulse and the vaccination rate closer to CDC. The three surveys also report that people without a four-year college degree are less likely to have been vaccinated compared to those with a degree (Table 2, Supplementary Table 1). If we assume that vaccination behaviours do not differ systematically between non-respondents and respondents within each demographic category, underrepresentation of less-vaccinated groups would contribute to the bias found here. However, this alone cannot explain the discrepancies in all the outcomes. Census Household Pulse weights on both race and education[4] and still overestimates vaccine uptake by over ten points in late May of 2021 (Fig. 1b).

Delphi–Facebook and Census Household Pulse may be unrepresentative with respect to political partisanship, which has been found to be correlated with vaccine behaviour[38] and with survey response[39], and thus may contribute to observed bias. However, neither Delphi–Facebook nor Census Household Pulse collects partisanship of respondents. US Census agencies cannot ask about political preference, and no unequivocal population benchmark for partisanship in the general adult population exists.

Rurality may also contribute to the errors, because it correlates with vaccine status[8] and home internet access[40]. Neither Census Household Pulse nor Delphi–Facebook weights on sub-state geography, which may mean that adults in more rural areas who are less likely to be vaccinated are also underrepresented in the two surveys, leading to overestimation of vaccine uptake.

Axios–Ipsos weights to metropolitan status and also recruits a fraction of its panellists from an 'offline' population of individuals without internet access. We find that dropping these offline respondents (n = 21, or 1% of the sample) in their 22 March 2021 wave increases Axios–Ipsos' overall estimate of the vaccination rate by 0.5 percentage points, thereby

increasing the total error (Extended Data Table 2). However, this offline population is too small to explain the entirety of the difference in accuracy between Axios–Ipsos and either Census Household Pulse (6 percentage points) or Delphi–Facebook (14 percentage points), in this time period.

Careful recruitment of panellists is at least as important as weighting. Weighting on observed covariates alone cannot explain or correct the discrepancies we observe. For example, reweighting Axios–Ipsos 22 March 2021 wave using only Delphi–Facebook's weighting variables (age group and gender) increased the error in their vaccination estimates by 1 percentage point, but this estimate with Axios–Ipsos data is still more accurate than that from Delphi–Facebook during the same period (Extended Data Table 2). The Axios–Ipsos estimate with Delphi–Facebook weighting overestimated vaccination by 2 percentage points, whereas Delphi–Facebook overestimated it by 11 percentage points.

The key implication is that there is no silver bullet: every small part of panel recruitment, sampling and weighting matters for controlling the data quality measured as the correlation between an outcome and response—what we call the ddc. In multi-stage sampling, which includes for example the selection of potential respondents followed by non-response, bias in even a single step can substantially affect the final result ('Population size in multi-stage sampling' in the Methods, Extended Data Table 3). A total quality control approach, inspired by the total survey error framework[41], is a better strategy than trying to prioritize some components over others to improve data quality. This emphasis is a reaffirmation of the best practice for survey research as advocated by the American Association for Public Opinion Research:[6] "The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise."[42]

Addressing common misperceptions

The three surveys discussed in this article demonstrate a seemingly paradoxical phenomenon—the two larger surveys that we studied are more statistically confident, but also more biased, than the smaller, more traditional Axios–Ipsos poll. These findings are paradoxical only when we fall into the trap of the intuition that estimation errors necessarily decrease in larger datasets[12].

A limitation of our vaccine uptake analysis is that we only examine ddc with respect to an outcome for which a benchmark is available: first-dose vaccine uptake. One might hope that surveys biased on vaccine uptake are not biased on other outcomes, for which there may not be benchmarks to reveal their biases. However, the absence of evidence of bias for the remaining outcomes is not evidence of its absence. In fact, mathematically, when a survey is found to be biased with respect to one variable, it implies that the entire survey fails to be statistically representative. The theory of survey sampling relies on statistical representativeness for all variables achieved through probabilistic sampling[43]. Indeed, Neyman's original introduction of probabilistic sampling showed the limits of purposive sampling, which attempted to achieve overall representativeness by enforcing it only on a set of variables[18,44].

In other words, when a survey loses its overall statistical representativeness (for example, through bias in coverage or non-response), which is difficult to repair (for example, by weighting or modelling on observable characteristics) and almost impossible to verify[45], researchers who wish to use the survey for scientific studies must supply other reasons to justify the reliability of their survey estimates, such as evidence about the independence between the variable of interest and the factors that are responsible for the unrepresentativeness. Furthermore, scientific journals that publish studies based on surveys that may be unrepresentative[17]—especially those with large sizes such as Delphi–Facebook (biased with respect to vaccination status (Fig. 1), race and education (Table 2))—need to ask for reasonable effort from the authors to address the unrepresentativeness.

Some may argue that bias is a necessary trade-off for having data that are sufficiently large for conducting highly granular analysis, such as county-level estimation of vaccine hesitancy[26]. Although high-resolution inference is important, we warn that this is a double-edged argument. A highly biased estimate with a misleadingly small confidence interval can do more damage than having no estimate at all. We further note that bias is not limited to population point estimates, but also affects estimates of changes over time (contrary to published guidance[3]). Both Delphi–Facebook and Census Household Pulse significantly overestimate the slope of vaccine uptake relative to that of the CDC benchmark (Fig. 1b).

The accuracy of our analysis does rely on the accuracy of the CDC's estimates of COVID vaccine uptake. However, if the selection bias in the CDC's benchmark is significant enough to alter our results, then that itself would be another example of the Big Data Paradox.

Discussion

This is not the first time that the Big Data Paradox has appeared: Google Trends predicted more than twice the number of influenza-like illnesses than the CDC in February 2013[46]. This analysis demonstrates that the Big Data Paradox applies not only to organically collected Big Data, like Google Trends, but also to surveys. Delphi–Facebook is "the largest public health survey ever conducted in the United States"[47]. The Census Household Pulse is conducted in collaboration between the US Census Bureau and eleven statistical government partners, all with enormous resources and survey expertise. Both studies take steps to mitigate selection bias, but substantially overestimate vaccine uptake. As we have shown, the effect of bias is magnified as relative sample size increases.

By contrast, Axios–Ipsos records only about 1,000 responses per wave, but makes additional efforts to prevent selection bias. Small surveys can be just as wrong as large surveys in expectation—of the three other small-to-medium online surveys additionally analysed, two also miss the CDC vaccination benchmark (Extended Data Fig. 5). The overall lesson is that investing in data quality (particularly during collection, but also in analysis) minimizes error more efficiently than does increasing data quantity. Of course, a sample size of 1,000 may be too small (that is, leading to unhelpfully large uncertainty intervals) for the kind of 50-state analyses made possible by big surveys. However, small-area methods that borrow information across subgroups[48] can perform better with higher-quality—albeit few—data, and whether that approach would outperform the large, biased surveys is an open question.

There are approaches to correct for these biases in both probability and nonprobability samples alike. For COVID-19 surveys in particular, since June 2021, the AP–NORC multimode panel has weighted their COVID-19 related surveys to the CDC benchmark, so that the weighted ddc for vaccine uptake is zero by design[49]. More generally, there is an extensive literature on approaches for making inferences from data collected from nonprobability samples[50–52]. Other promising approaches include integrating surveys of varying quality[53,54], and leveraging the estimated ddc in one outcome to correct bias in others under several scenarios (Supplementary Information D).

Although more needs to be done to fully examine the nuances of large surveys, organically collected administrative datasets and social media data, we hope that this comparative study of ddc highlights the concerning implications of the Big Data Paradox—how large sample sizes magnify the effect of seemingly small defects in data collection, which leads to overconfidence in incorrect inferences.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04198-4.

1. Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12**, 685–726 (2018).
2. Barkay, N. et al. Weights and methodology brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in partnership with Facebook. Preprint at https://arxiv.org/abs/2009.14675 (2020).
3. Kreuter, F. et al. Partnering with Facebook on a university-based rapid turn-around global survey. *Surv. Res. Methods* **14**, 159–163 (2020).
4. Fields, J. F. et al. *Design and Operation of the 2020 Household Pulse Survey* (U.S. Census Bureau, 2020).
5. Jackson, C., Newall, M. & Yi, J. Axios Ipsos Coronavirus Index (2021).
6. American Association for Public Opinion Research (AAPOR). Best practices for survey research. https://www.aapor.org/Standards-Ethics/Best-Practices.aspx (2021).
7. Hastak, M., Mazis, M. B. & Morris, L. A. The role of consumer surveys in public policy decision making. *J. Public Policy Mark.* **20**, 170–185 (2001).
8. B. P. Murthy, et al. Disparities in COVID-19 vaccination coverage between urban and rural counties: United States, December 14, 2020–April 10, 2021. *MMWR Morb. Mortal. Wkly Rep.* **70**, 759–764 (2021).
9. Arrieta, A., Gakidou, E., Larson, H., Mullany, E. & Troeger, C. Through understanding and empathy, we can convince women to get the COVID-19 vaccine. *Think Global Health* https://www.thinkglobalhealth.org/article/through-understanding-and-empathy-we-can-convince-women-get-covid-19-vaccine (2021).
10. Japec, L. et al. *AAPOR Report on Big Data* (American Association of Public Opinion Researchers, 2015).
11. Reinhart, A., Kim, E., Garcia, A. & LaRocca, S. Using the COVID-19 Symptom Survey to track vaccination uptake and sentiment in the United States. *CMU Delphi Group* https://delphi.cmu.edu/blog/2021/01/28/using-the-covid-19-symptom-survey-to-track-vaccination-uptake-and-sentiment-in-the-united-states (2021).
12. Mayer-Schönberger, V. & Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt, 2013).
13. CDC. Trends in number of COVID-19 vaccinations (2021).
14. Nguyen, K. H. et al. Comparison of COVID-19 vaccination coverage estimates from the Household Pulse Survey, Omnibus Panel Surveys, and COVID-19 vaccine administration data, United States, March 2021. *CDC AdultVaxView* https://www.cdc.gov/vaccines/imz-managers/coverage/adultvaxview/pubs-resources/covid19-coverage-estimates-comparison.html (2021).
15. Santibanez, T. A. et al. Sociodemographic factors associated with receipt of COVID-19 vaccination and intent to definitely get vaccinated, adults aged 18 years or above—Household Pulse Survey, United States, April 28–May 10, 2021. *CDC AdultVaxView* https://www.cdc.gov/vaccines/imz-managers/coverage/adultvaxview/pubs-resources/sociodemographic-factors-covid19-vaccination.html (2021).
16. Kruskal, W. & Mosteller, F. Representative sampling, I: Non-scientific literature. *Int. Stat. Rev.* **47**,13–24 (1979).
17. Kruskal, W. & Mosteller, F. Representative sampling, II: Scientific literature, excluding statistics. *Int. Stat. Rev.* **47**, 111–127 (1979).
18. Kruskal, W. & Mosteller, F. Representative sampling, III: The current statistical literature. *Int. Stat. Rev*. **47**, 245–265 (1979).
19. Kruskal, W. & Mosteller, F. Representative sampling, IV: The history of the concept in statistics, 1895–1939. *Int. Stat. Rev.* **48**, 169–195 (1980).
20. American Association for Public Opinion Research (AAPOR). Margin of sampling error/credibility interval. https://www.aapor.org/Education-Resources/Election-Polling-Resources/Margin-of-Sampling-Error-Credibility-Interval.aspx (2021).
21. The Delphi Group at Carnegie Mellon University in partnership with Facebook. *Topline Report on COVID-19 Vaccination in the United States* (2021).
22. Haas, E. J. et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *Lancet* **397**, 1819–1829 (2021).
23. Kish, L. *Survey Sampling* (Wiley, 1965).
24. Institute for Health Metrics and Evaluation (IHME). COVID-19 vaccine hesitancy. https://vaccine-hesitancy.healthdata.org/ (2021).
25. King, W. C., Rubinstein, M., Reinhart, A. & Mejia, R. J. Time trends and factors related to COVID-19 vaccine hesitancy from january-may 2021 among US adults: findings from a large-scale national survey. Preprint at https://doi.org/10.1101/2021.07.20.21260795 (2021).
26. CDC. Estimates of vaccine hesitancy for COVID-19 (2021).
27. Groves, R. M. et al. *Survey Methodology* Vol. 561 (Wiley, 2011).
28. Dempsey, W. The hypothesis of testing: paradoxes arising out of reported coronavirus case-counts. Preprint at https://arxiv.org/abs/2005.10425 (2020).
29. Isakov, M. & Kuriwaki, S. Towards principled unskewing: viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review* **2** https://doi.org/10.1162/99608f92.86a46f38 (2020).
30. Hartley, H. O. & Ross, A. Unbiased ratio estimators. *Nature* **174**, 270–271 (1954).
31. Tiu, A., Susswein, Z., Merritt, A. & Bansal, S. Characterizing the spatiotemporal heterogeneity of the COVID-19 vaccination landscape. Preprint at https://doi.org/10.1101/2021.10.04.21263345 (2021).
32. Groen, J. Sources of error in survey and administrative data: the importance of reporting procedures. *J. Off. Stat.* **28**, 173–198 (2012).
33. Tu, X. M., Meng, X.-L. & Pagano, M. The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Am. Stat. Assoc.* **88**, 26–36 (1993).
34. Barnes, O. & Burn-Murdoch, J. COVID response hampered by population data glitches. *Financial Times* (11 October 2021).
35. Kennedy, C. et al. Evaluating online nonprobability surveys. *Pew Research Center* https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/ (2016).
36. Kennedy, C. et al. An evaluation of the 2016 election polls in the United States. *Public Opin. Q.* **82**, 1–33 (2018).
37. Auxier, B. & Anderson, M. Social media use in 2021. *Pew Research Center* https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ (2021).
38. Gadarian, S. K., Goodman, S. W. & Pepinsky, T. B. Partisanship, health behavior, and policy attitudes in the early stages of the COVID-19 pandemic. *PLoS ONE* **16**, e0249596 (2021).
39. Mercer, A., Lau, A. & Kennedy, C. For weighting online opt-in samples, what matters most? *Pew Research Center* https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/ (2018).
40. Ryan, C. *Computer and Internet Use in the United States: 2016*. American Community Survey Report No. ACS-39 (U.S. Census Bureau, 2017).
41. Biemer, P. P. & Lyberg, L. E. *Introduction to Survey Quality* (Wiley, 2003).
42. Scheuren, F. *What is a Survey?* (American Statistical Association, 2004).
43. Sukhatme, P. V. *Sampling Theory of Surveys with Applications* (1954).
44. Neyman, J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.* **97**, 558–625 (1934).
45. Groves, R. M. Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Q.* **70**, 646–675 (2006).
46. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
47. Salomon, J. A. et al. The US COVID-19 Trends and Impact Survey, 2020-2021: continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing and vaccination. Preprint at https://doi.org/10.1101/2021.07.24.21261076 (2021).
48. Park, D. K., Gelman, A. & Bafumi, J. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Polit. Anal.* **12**, 375–385 (2004).
49. Associated Press–NORC Center for Public Affairs Research. The June 2021 AP–NORC center poll (2021).
50. Wang, W., Rothschild, D., Goel, S. & Gelman, A. Forecasting elections with non-representative polls. *Int. J. Forecast.* **31**, 980–991 (2015).
51. Elliott, M. R. & Valliant, R. Inference for nonprobability samples. *Stat. Sci.* **32**, 249–264 (2017).
52. Little, R. J., West, B. T., Boonstra, P. S. & Hu, J. Measures of the degree of departure from ignorable sample selection. *J. Surv. Stat. Methodol.* **8**, 932–964 (2020).
53. Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A. & Blom, A. G. Integrating probability and nonprobability samples for survey inference. *J. Surv. Stat. Methodol.***8**, 120–147 (2020).
54. Yang, S., Kim, J. K. & Song, R. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *J. R. Stat. Soc. B* **82**, 445–465 (2020).

Methods

Calculation and interpretation of ddc

The mathematical expression for equation (1) is given here for completeness:

$$\bar{Y}_n - \bar{Y}_N = \hat{\rho}_{Y,R} \times \sqrt{\frac{N-n}{n}} \times \sigma_Y \quad (2)$$

The first factor $\hat{\rho}_{Y,R}$ is called the data defect correlation (ddc)[1]. It is a measure of data quality represented by the correlation between the recording indicator R ($R = 1$ if an answer is recorded and $R = 0$ otherwise) and its value, Y. Given a benchmark, the ddc $\hat{\rho}_{Y,R}$ can be calculated by substituting known quantities into equation (2). In the case of a single survey wave of a COVID-19 survey, n is the sample size of the survey wave, N is the population size of US adults from US Census estimates[55], $\bar{Y}_n$ is the survey estimate of vaccine uptake and $\bar{Y}_N$ is the estimate of vaccine uptake for the corresponding period taken from the CDC's report of the cumulative count of first doses administered to US adults[8,13]. We calculate $\sigma_Y = \sqrt{\bar{Y}_N(1-\bar{Y}_N)}$ because Y is binary (but equation (2) is not restricted to binary Y).

We calculate $\hat{\rho}_{Y,R}$ by using total error $\bar{Y}_n - \bar{Y}_N$, which captures not only selection bias but also any measurement bias (for example, from question wording). However, with this calculation method, $\hat{\rho}_{Y,R}$ lacks the direct interpretation as a correlation between Y and R, and instead becomes a more general index of data quality directly related to classical design effects (see 'Bias-adjusted effective sample size').

It is important to point out that the increase in ddc does not necessarily imply that the response mechanisms for Delphi–Facebook and Census Household Pulse have changed over time. The correlation between a changing outcome and a steady response mechanism could change over time, hence changing the value of ddc. For example, as more individuals become vaccinated, and vaccination status is driven by individual behaviour rather than eligibility, the correlation between vaccination status and propensity to respond could increase even if the propensity to respond for a given individual is constant. This would lead to large values of ddc over time, reflecting the increased impact of the same response mechanism.

Error decomposition with survey weights

The data quality framework given by equations (1) and (2) is a special case of a more general framework for assessing the actual error of a weighted estimator $\bar{Y}_w = \frac{\sum_i w_i R_i Y_i}{\sum_i w_i R_i}$, where w_i is the survey weight assigned to individual i. It is shown in Meng[1] that

$$\bar{Y}_{\mathrm{w}} - \bar{Y}_N = \hat{\rho}_{Y,R_{\mathrm{w}}} \times \sqrt{\frac{N-n_{\mathrm{w}}}{n_{\mathrm{w}}}} \times \sigma_Y, \quad (3)$$

where $\hat{\rho}_{Y,R_{\mathrm{w}}} = \mathrm{Corr}(Y, R_{\mathrm{w}})$ is the finite population correlation between Y_i and $R_{\mathrm{w},i} = w_i R_i$ (over $i = 1, \ldots, N$). The 'hat' on ρ reminds us that this correlation depends on the specific realization of $\{R_i, i = 1, \ldots, N\}$. The term n_{w} is the classical 'effective sample size' due to weighting[23]; that is, $n_{\mathrm{w}} = \frac{n}{(1+\mathrm{CV}_{\mathrm{w}}^2)}$, where CV_{w} is the coefficient of variation of the weights for all individuals in the observed sample, that is, the standard deviation of weights normalized by their mean. It is common for surveys to rescale their weights to have mean 1, in which case CV_w^2 is simply the sample variance of W.

When all weights are the same, equation (3) reduces to equation (2). In other words, the ddc term $\hat{\rho}_{Y,R_{\mathrm{w}}}$ now also takes into account the effect of the weights as a means to combat the selection bias represented by the recording indicator R. Intuitively, if $\hat{\rho}_{Y,R} = \mathrm{Corr}(Y, R)$ is high (in magnitude), then some Y_i's have a higher chance of entering our dataset than others, thus leading to a sample average that is a biased estimator for the population average. Incorporating appropriate weights can reduce $\hat{\rho}_{Y,R}$ to $\hat{\rho}_{Y,R_{\mathrm{w}}}$, with the aim of reducing the effect of the selection bias. However, this reduction alone may not be sufficient to improve the accuracy of $\bar{Y}_w$ because the use of weight necessarily reduces the sampling fraction $f = \frac{n}{N}$ to $f_{\mathrm{w}} = \frac{n_{\mathrm{w}}}{N}$ as well, as $n_{\mathrm{w}} < n$. Equation (3) precisely describes this trade-off, providing a formula to assess when the reduction of ddc is significant to outweigh the reduction of the effective sample size.

Measuring the correlation between Y and R is not a new idea in survey statistics (though note that ddc is the population correlation between Y and R, not the sample correlation), nor is the observation that as sample size increases, error is dominated by bias instead of variance[56,57]. The new insight is that ddc is a general metric to index the lack of representativeness of the data we observe, regardless of whether or not the sample is obtained through a probabilistic scheme, or weighted to mimic a probabilistic sample. As discussed in 'Addressing common misperceptions' in the main text, any single ddc deviating from what is expected under representative sampling (for example, probabilistic sampling) is sufficient to establish that the sample is not representative (but the converse is not true). Furthermore, the ddc framework refutes the common belief that increasing sample size necessarily improves statistical estimation[1,58].

Bias-adjusted effective sample size

By matching the mean-squared error of $\bar{Y}_w$ with the variance of the sample average from simple random sampling, Meng[1] derives the following formula for calculating a bias-adjusted effective sample size, or n_{eff}:

$$n_{\mathrm{eff}} = \frac{n_{\mathrm{w}}}{N - n_{\mathrm{w}}} \times \frac{1}{E\left[\hat{\rho}^2_{Y,R_{\mathrm{w}}}\right]}$$

Given an estimator $\bar{Y}_w$ with expected total MSE T due to data defect, sampling variability and weighting, this quantity n_{eff} represents the size of a simple random sample such that its mean $\bar{Y}_N$, as an estimator for the same population mean $\bar{Y}_N$, would have the identical MSE T. The term $E[\hat{\rho}^2_{Y,R_{\mathrm{w}}}]$ represents the amount of selection bias (squared) expected on average from a particular recording mechanism R and a chosen weighting scheme.

For each survey wave, we use $\hat{\rho}^2_{Y,R_{\mathrm{w}}}$ to approximate $E[\hat{\rho}^2_{Y,R_{\mathrm{w}}}]$. This estimation is unbiased by design, as we use an estimator to estimate its expectation. Therefore, the only source of error is the sampling variation, which is typically negligible for large surveys such as Delphi–Facebook and the Census Household Pulse. This estimation error may have more impact for smaller surveys such as the Axios–Ipsos survey, an issue that we will investigate in subsequent work.

We compute $\hat{\rho}_{Y,R_{\mathrm{w}}}$ by using the benchmark $\bar{Y}_N$, namely, by solving equation (3) for $\hat{\rho}_{Y,R_{\mathrm{w}}}$,

$$\hat{\rho}_{Y,R_{\mathrm{w}}} = \frac{Z_{\mathrm{w}}}{\sqrt{N}}, \text{ where } Z_{\mathrm{w}} = \frac{\bar{Y}_{\mathrm{w}} - \bar{Y}_N}{\sqrt{\frac{1-f_{\mathrm{w}}}{n_{\mathrm{w}}}}\sigma_Y}$$

We introduce this notation Z_{w} because it is the quantity that determines the well-known survey efficiency measure, the so-called 'design effect', which is the variance of Z_{w} for a probabilistic sampling design[23] (when we assume the weights are fixed). For the more general setting in which $\bar{Y}_w$ may be biased, we replace the variance by MSE, and hence the bias-adjusted design effect $D_e = E[Z^2_{\mathrm{w}}]$, which is the MSE relative to the benchmark measured in the unit of the variance of an average from a simple random sample of size n_{w}. Hence $D_I \equiv E[\hat{\rho}^2_{Y,R_{\mathrm{w}}}]$, which was termed as 'data defect index'[1], is simply the bias-adjusted design effect per unit, because $D_I = \frac{D_e}{N}$.

Furthermore, because Z_{w} is the standardized actual error, it captures any kind of error inherited in $\bar{Y}_w$. This observation is important because when Y is subject to measurement errors, $\frac{Z_{\mathrm{w}}}{\sqrt{N}}$ no longer has the simple interpretation as a correlation. But because we estimate D_I by $\frac{Z^2_w}{N}$

directly, our effective sample size calculation is still valid even when equation (3) does not hold.

Asymptotic behaviour of ddc

As shown in Meng[1], for any probabilistic sample without selection biases, the ddc is on the order of $\frac{1}{\sqrt{N}}$. Hence the magnitude of $\hat{\rho}_{Y,R}$ (or $\hat{\rho}_{Y,R_W}$) is small enough to cancel out the effect of $\sqrt{N-n}$ (or $\sqrt{N-n_w}$) in the data scarcity term on the actual error, as seen in equation (2) (or equation (3)). However, when a sample is unrepresentative; for example, when those with $Y=1$ are more likely to enter the dataset than those with $Y=0$, then $\hat{\rho}_{Y,R}$ can far exceed $\frac{1}{\sqrt{N}}$ in magnitude. In this case, error will increase with $\sqrt{N}$ for a fixed ddc and growing population size N (equation (2)). This result may be counterintuitive in the traditional survey statistics framework, which often considers how error changes as sample size n grows. The ddc framework considers a more general set-up, taking into account individual response behaviour, including its effect on sample size itself.

As an example of how response behaviour can shape both total error and the number of respondents n, suppose individual response behaviour is captured by a logistic regression model

$$\text{logit}[\Pr(R=1|Y)] = \alpha + \beta Y. \quad (4)$$

This is a model for a response propensity score. Its value is determined by α, which drives the overall sampling fraction $f=\frac{n}{N}$, and by β, which controls how strongly Y influences whether a participant will respond or not.

In this logit response model, when $\beta \neq 0$, $\hat{\rho}_{Y,R}$ is determined by individual behaviour, not by population size N. In Supplementary Information B.1, we prove that ddc cannot vanish as N grows, nor can the observed sample size n ever approach 0 or N for a given set of (finite and plausible) values of $\{\alpha, \beta\}$, because there will always be a non-trivial percentage of non-respondents. For example, an f of 0.01 can be obtained under this model for either $\alpha=-0.46$, $\beta=0$ (no influence of individual behaviour on response propensity), or for $\alpha=-3.9$, $\beta=-4.84$. However, despite the same f, the implied ddc and consequently the MSE will differ. For example, the MSE for the former (no correlation with Y) is 0.0004, whereas the MSE for the latter (a −4.84 coefficient on Y) is 0.242, over 600 times larger.

See Supplementary Information B.2 for the connection between ddc and a well-studied non-response model from econometrics, the Heckman selection model[59].

Population size in multi-stage sampling

We have shown that the asymptotic behaviour of error depends on whether the data collection process is driven by individual response behaviour or by survey design. The reality is often a mix of both. Consequently, the relevant 'population size' N depends on when and where the representativeness of the sample is destroyed; that is, when the individual response behaviours come into play. Real-world surveys that are as complex as the three surveys we analyse here have multiple stages of sample selection.

Extended Data Table 3 takes as an example the sampling stages of the Census Household Pulse, which has the most extensive set of documentation among the three surveys we analyse. As we have summarized (Table 1, Extended Data Table 1), the Census Household Pulse (1) first defines the sampling frame as the reachable subset of the MAF, (2) takes a random sample of that population to prompt (send a survey questionnaire) and (3) waits for individuals to respond to that survey. Each of these stages reduces the desired data size, and the corresponding population size is the intended sample size from the prior stage (in notation, $N_s = n_{s-1}$, for $s=2, 3$). For example, in stage 3, the population size N_3 is the size of the intended sample size n_2 from the second stage (random sample of the outreach list), because only the sampled individuals have a chance to respond.

Although all stages contribute to the overall ddc, the stage that dominates is the first stage at which the representativeness of our sample is destroyed—the size of which will be labelled as the dominating population size (dps)—when the relevant population size decreases markedly at each step. However, we must bear in mind that dps refers to the worst-case scenario, when biases accumulate, instead of (accidentally) cancelling each other out.

For example, if the 20% of the MAFs excluded from the Census Household Pulse sampling frame (because they had no cell phone or email contact information) is not representative of the US adult population, then the dps is N_1, or 255 million adults contained in 144 million households. Then the increase in bias for given ddc is driven by the rate of $\sqrt{N_1}$ where $N_1 = 2.55 \times 10^8$ and is large indeed (with $\sqrt{2.5 \times 10^8} \approx 15{,}000$). By contrast, if the the sampling frame is representative of the target population and the outreach list is representative of the frame (and hence representative of the US adult population) but there is non-response bias, then dps is $N_3 = 10^6$ and the impact of ddc is amplified by the square root of that number ($\sqrt{10^6} = 1{,}000$). By contrast, Axios–Ipsos reports a response rate of about 50%, and obtains a sample of $n = 1{,}000$, so the dps could be as small as $N_3 = 2{,}000$ (with $\sqrt{2{,}000} \approx 45$).

This decomposition is why our comparison of the surveys is consistent with the 'Law of Large Populations'[1] (estimation error increases with $\sqrt{N}$), even though all three surveys ultimately target the same US adult population. Given our existing knowledge about online–offline populations[40] and our analysis of Axios–Ipsos' small 'offline' population, Census Household Pulse may suffer from unrepresentativeness at Stage 1 of Extended Data Table 3, where $N = 255$ million, and Delphi–Facebook may suffer from unrepresentativeness at the initial stage of starting from the Facebook user base. By contrast, the main source of unrepresentativeness for Axios–Ipsos may be at a later stage at which the relevant population size is orders of magnitude smaller.

CDC estimates of vaccination rates

Our analysis of the nationwide vaccination rate covers the period between 9 January 2021 and 19 May 2021. We used CDC's vaccination statistics published on their data tracker as of 26 May 2021. This dataset is a time series of counts of 1st dose vaccinations for every day in our time period, reported for all ages and disaggregated by age group.

This CDC time series obtained on 26 May 2021 included retroactive updates to dates covering our entire study period, as does each daily update provided by the CDC daily update. For example, the CDC benchmark we use for March 2021 is not only the vaccination counts originally reported in March but also includes the delayed reporting for March that the CDC became aware of by 26 May 2021. Analyzing several snapshots before 26 May 2021, we find that these retroactive updates 40 days out could change the initial estimate by about 5% (Extended Data Fig. 3), hence informing our sensitivity analysis of +/− 5% and 10% benchmark imprecision.

To match the sampling frame of the surveys we analyze, US adults 18 years and older, we must restrict the CDC vaccination counts to those administered to those adults. However, because of the different way states and jurisdiction report their vaccination statistics, the CDC did not possess age-coded counts for some jurisdictions, such as Texas, at the time of our study. The number of vaccinations with missing age data reached about 10 percent of the total US vaccinations at its peak at the time of our study. We therefore assume that the day by day fraction of adults among individuals for whom age is reported as missing is equal to the fraction of adults among individuals with age reported. Because minors became eligible for vaccinations only towards the end of our study period, the fraction of adults in data reporting age never falls below 97%.

Additional survey methodology

The Census Household Pulse and Delphi–Facebook surveys are the first of their kind for each organization, whereas Ipsos has maintained their online panel for 12 years.

Question wording

All three surveys ask whether respondents have received a COVID-19 vaccine (Extended Data Table 1). Delphi–Facebook and Census Household Pulse ask similar questions ("Have you had/received a COVID-19 vaccination/vaccine?"). Axios–Ipsos asks "Do you personally know anyone who has already received the COVID-19 vaccine?", and respondents are given response options including "Yes, I have received the vaccine." The Axios–Ipsos question wording might pressure respondents to conform to their communities' modal behaviour and thus misreport their true vaccination status, or may induce acquiescence bias from the multiple 'yes' options presented[60]. This pressure may exist both in high- and low-vaccination communities, so its net effect on Axios–Ipsos' results is unclear. Nonetheless, Axios–Ipsos' question wording does differ from that of the other two surveys, and may contribute to the observed differences in estimates of vaccine uptake across surveys.

Population of interest

All three surveys target the US adult population, but with different sampling and weighting schemes. Census Household Pulse sets the denominator of their percentages as the household civilian, non-institutionalized population in the United States of 18 years of age or older, excluding Puerto Rico or the island areas. Axios–Ipsos designs samples to be representative of the US general adult population of 18 or older. For Delphi–Facebook, the US target population reported in weekly contingency tables is the US adult population, excluding Puerto Rico and other US territories. For the CDC Benchmark, we define the denominator as the US 18+ population, excluding Puerto Rico and other US territories. To estimate the size of the total US population, we use the US Census Bureau Annual Estimates of the Resident Population for the United States and Puerto Rico, 2019[55]. This is also what the CDC uses as the denominator in calculating rates and percentages of the US population[60].

Axios–Ipsos and Delphi–Facebook generate target distributions of the US adult population using the Current Population Survey (CPS), March Supplement, from 2019 and 2018, respectively. Census Household Pulse uses a combination of 2018 1-year American Community Survey (ACS) estimates and the Census Bureau's Population Estimates Program (PEP) from July 2020. Both the CPS and ACS are well-established large surveys by the Census and the choice between them is largely inconsequential.

Axios–Ipsos data

The Axios–Ipsos Coronavirus tracker is an ongoing, bi-weekly tracker intended to measure attitudes towards COVID-19 of adults in the US. The tracker has been running since 13 March 2020 and has released results from 45 waves as of 28 May 2021. Each wave generally runs over a period of 4 days. The Axios–Ipsos data used in this analysis were scraped from the topline PDF reports released on the Ipsos website[5]. The PDF reports also contain Ipsos' design effects, which we have confirmed are calculated as 1 plus the variance of the (scaled) weights.

Census Household Pulse data

The Census Household Pulse is an experimental product of the US Census Bureau in collaboration with eleven other federal statistical agencies. We use the point estimates presented in Data Tables, as well as the standard errors calculated by the Census Bureau using replicate weights. The design effects are not reported, however we can calculate it as $1+\mathrm{CV}_w^2$, where CV_w is the coefficient of variation of the individual-level weights included in the microdata[23].

Delphi–Facebook COVID symptom survey

The Delphi–Facebook COVID symptom survey is an ongoing survey collaboration between Facebook, the Delphi Group at Carnegie Mellon University (CMU), and the University of Maryland[2]. The survey is intended to track COVID-like symptoms over time in the US and in over 200 countries. We use only the US data in this analysis. The study recruits respondents using daily stratified random samples recruiting a cross-section of Facebook active users. New respondents are obtained each day, and aggregates are reported publicly on weekly and monthly frequencies. The Delphi–Facebook data used here were downloaded directly from CMU's repository for weekly contingency tables with point estimates and standard errors.

Ethical compliance

According to HRA decision tools (http://www.hra-decisiontools.org.uk/research/), our study is considered Research, and according to the NHS REC review tool (http://www.hra-decisiontools.org.uk/ethics/), we do not need NHS Research Ethics Committee (REC) review, as we used only (1) publicly available, (2) anonymized and (3) aggregated data outside of clinical settings.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw data have been deposited in the Harvard Dataverse, at https://doi.org/10.7910/DVN/GKBUUK. Data were collected from publicly available repositories of survey data by downloading the data directly or using APIs.

Code availability

Code to replicate the findings is available in the repository https://github.com/vcbradley/ddc-vaccine-US. The main decomposition of the ddc is available on the package 'ddi' from the Comprehensive R Archive Network (CRAN).

55. U.S. Census Bureau. Methodology for the United States population estimates: Vintage 2019.
56. Bethlehem, J. in *Survey Nonresponse* (eds Groves, R. M. et al.) 275–288 (Wiley, 2002).
57. Meng, X.-L. in *Past, Present, and Future of Statistical Science* (eds Lin, X. et al.) 537–562 (CRC Press, 2014).
58. Meng, X.-L. & Xie, X. I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econom. Rev.* **33**, 218–250 (2014).
59. Heckman, J. J. Sample selection bias as a specification error. *Econometrica* 153–161 (1979).
60. CDC. Reporting COVID-19 vaccination demographic data. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/demographics-vaccination-data.html (2021).

Acknowledgements We thank F. Kreuter, A. Reinhart and the Delphi Group at CMU; Facebook's Demography and Survey Science group; F. Barlas, C. Jackson, C. Morris, M. Newall and the Public Affairs team at Ipsos; and J. Fields and J. Hunter Childs at the US Census Bureau for productive conversations about their surveys. We further thank the Delphi Group at CMU for their help in computing weekly design effects for their survey; the Ipsos team for providing data on their 'offline' respondents; the CDC for responding to our questions; S. Paddock, other participants at the JPSM 2021 lecture (delivered by X.-L.M.) and S. Finch for providing comments; A. Edwards-Levy for inspiring our interest in this topic; and R. Born for suggesting more intuitive terms used in equation (1). V.C.B. is funded by the University of Oxford's Clarendon Fund and the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). X.-L.M. acknowledges partial financial support by NSF. S.F. acknowledges the support of the EPSRC (EP/V002910/1).

Author contributions V.C.B. and S.F. conceived and formulated the research questions. V.C.B. and S.K. contributed equally to data analysis, writing and visualization. X.-L.M. conceived and formulated the methodology. All authors contributed to methodology, writing, visualization, editing and data analysis. S.F. supervised the work.

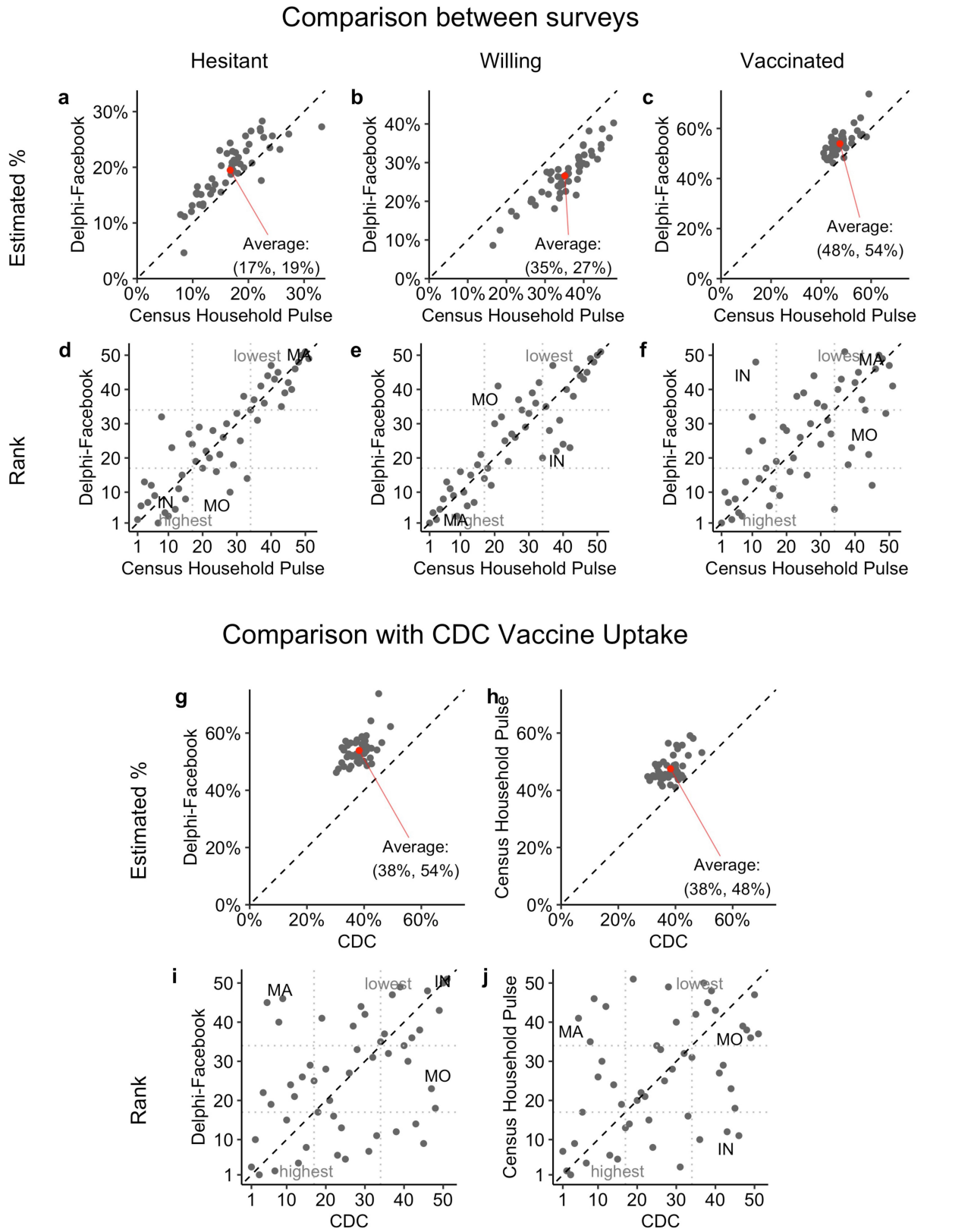
Competing interests The authors declare no competing interests.

Additional information

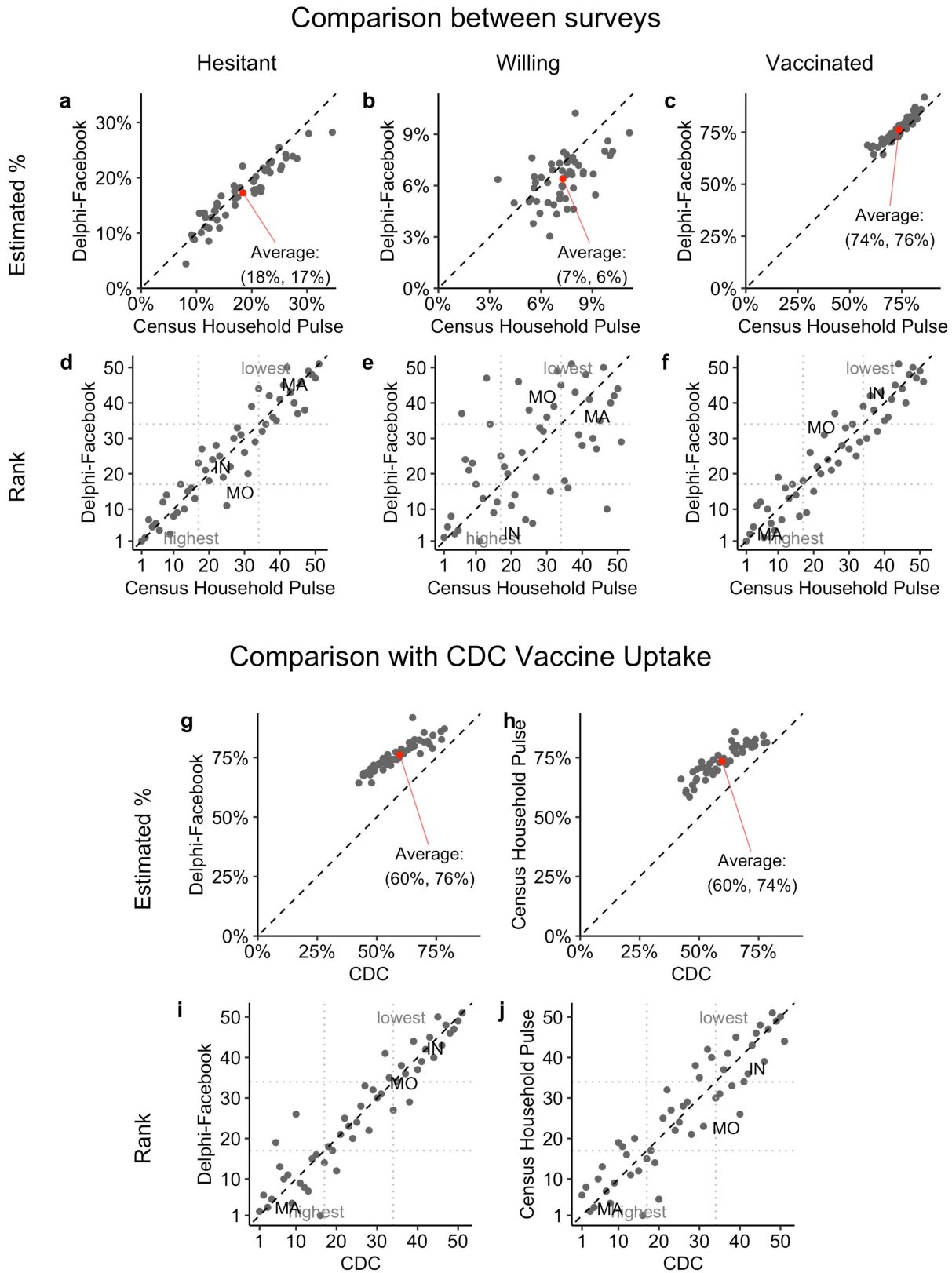
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04198-4.

Correspondence and requests for materials should be addressed to Seth Flaxman.

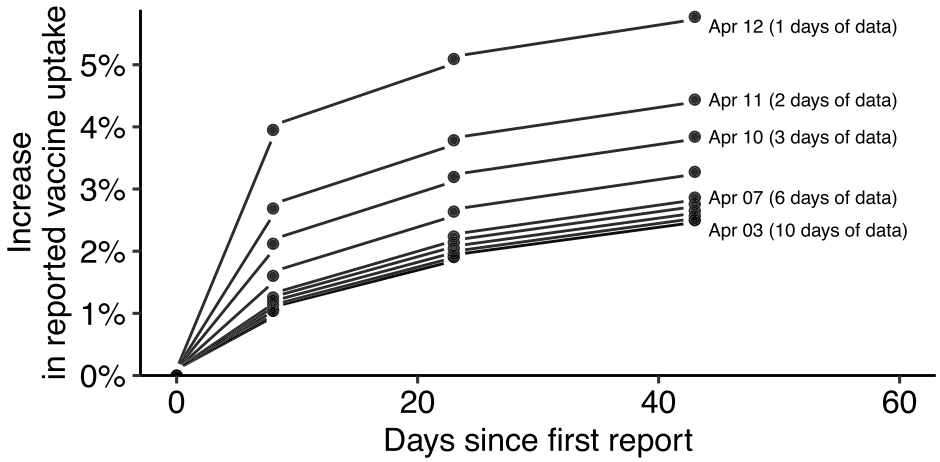
Peer review information *Nature* thanks Michael Elliot, Alex Reinhart and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



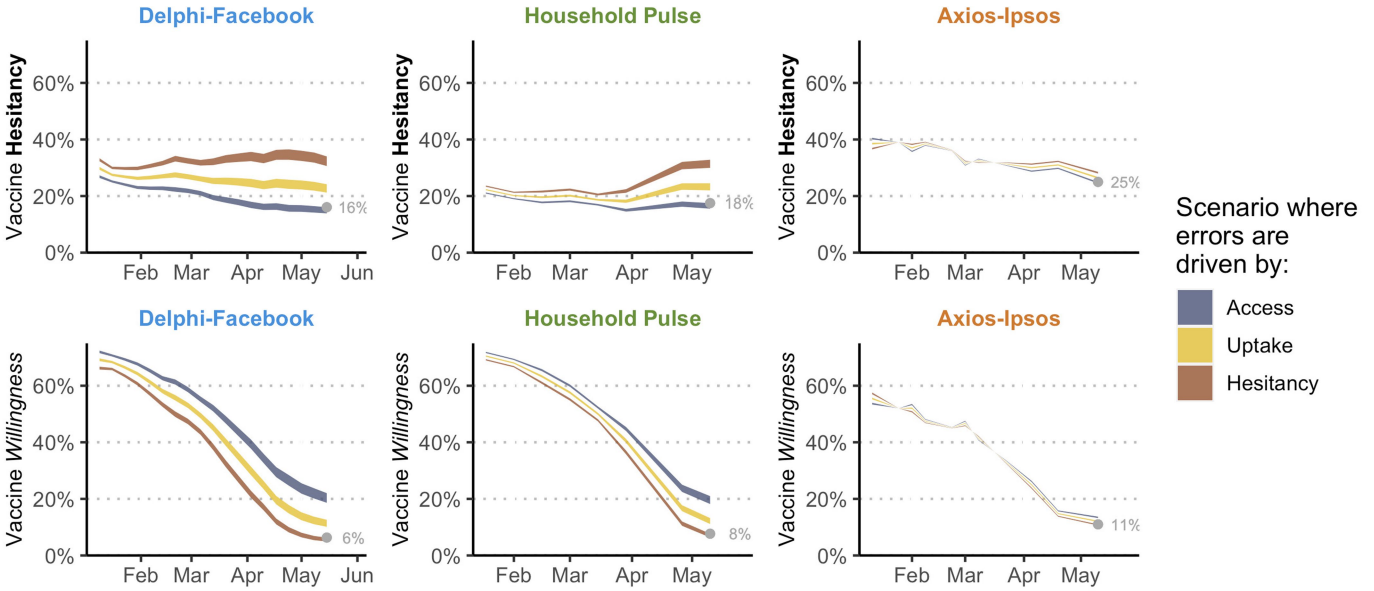
Extended Data Fig. 1 | Comparisons of state-level vaccine uptake, hesitancy and willingness across surveys and the CDC for March 2021. Comparison of Delphi-Facebook and Census Household Pulse's state-level point estimates (**a**–**c**) and rankings (**d**–**f**) for vaccine hesitancy, willingness and uptake Dotted black lines show agreement and red points show the average of 50 states. During our study period, the CDC published daily reports of the cumulative number of vaccinations by state that had occurred up to a certain date. Due to reporting delays, these may be an underestimate, but retroactively updated data was not available to us. **g**–**j** compare state-level point estimates and rankings for the same survey waves to CDC benchmark estimates from 31 March 2021. The Delphi–Facebook data are from the week ending 27 March 2021 and the Census Household Pulse is the wave ending 29 March 2021. See Extended Data Fig. 3 for details on the degree of retroactive updates we could expect, and Supplementary Information A.2 for details.



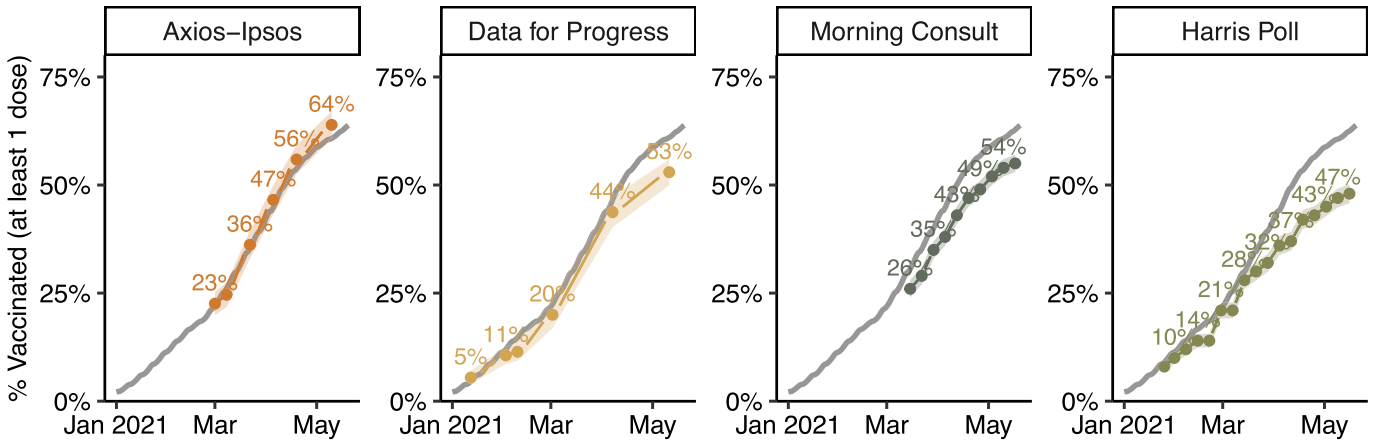
Extended Data Fig. 2 | Comparisons of state-level vaccine uptake, hesitancy and willingness across surveys and the CDC for May 2021. Comparison of Delphi-Facebook and Census Household Pulse's state-level point estimates (**a**–**c**) and rankings (**d**–**f**) for vaccine hesitancy, willingness and uptake. Dotted black lines show agreement and red points show the average of 50 states. During our study period, the CDC published daily reports of the cumulative number of vaccinations by state that had occurred up to a certain date. Due to reporting delays, these may be an underestimate, but retroactively updated data was not available to us. **g**–**j** compare state-level point estimates and rankings for the same survey waves to CDC benchmark estimates from 15 May 2021. The Delphi–Facebook data are from the wave week ending 8 May 2021 and the Census Household Pulse is the wave ending 10 May 2021. See Extended Data Fig. 3 for details on the degree of retroactive updates we could expect, and Supplementary Information A.2 for details.



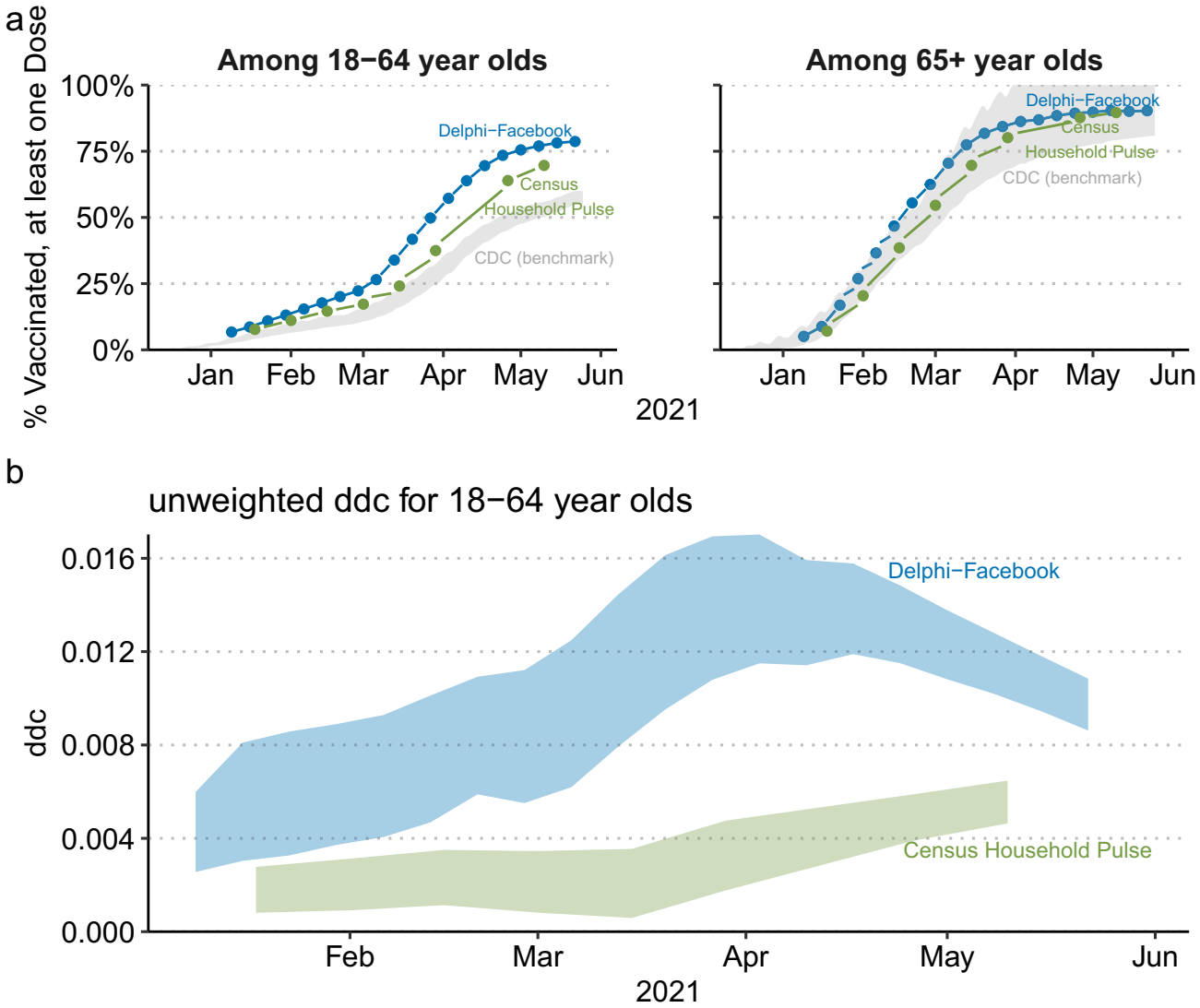
Extended Data Fig. 3 | Retroactive adjustment of CDC vaccine uptake figures for 3–12 April 2021, over the 45 days from 12 April. Increase is shown as a percentage of the vaccine uptake reported on 12 April. Most of the retroactive increases in reported estimates appear to occur in the first 10 days after an estimate is first reported. By about 40 days after the initial estimates for a particular day are reported, the upward adjustment plateaus at around 5–6% of the initial estimate. We use this analysis to guide the choice of 5% and 10% threshold for the possible imprecision in the CDC benchmark when computing Benchmark Imprecision (BI) intervals.



Extended Data Fig. 4 | Revised estimates of hesitancy and willingness after accounting for survey errors for vaccination uptake. The grey point shows the reported value at the last point of the time series. Each line shows a different scenario for what might be driving the error in uptake estimate, derived using hypothetical ddc values for willingness and hesitancy based on the observed ddc value for uptake. Access scenario: willingness suffers from at least as much, if not more, bias than uptake. Hesitancy scenario: hesitancy suffers from at least as much, if not more, bias than uptake. Uptake scenario: the error is split roughly equally between hesitancy and willingness. See Supplementary Information D for more details.



Extended Data Fig. 5 | Vaccination rates compared with CDC benchmark for four online polls. Ribbons indicate traditional 95% confidence intervals, which are twice the standard error reported by the poll. Grey line is the CDC benchmark. Data for Progress asks "As of today, have you been vaccinated for Covid-19?"; Morning Consult asks "Have you gotten the vaccine, or not?"; Harris Poll asks "Which of the following best describes your mindset when it comes to getting the COVID-19 vaccine when it becomes available to you?". YouGov surveys are not analysed because they explicitly examined how their surveys tracked CDC vaccine uptake. See Supplementary Information C.3 for the sampling methodology of each survey and discussion of differences.



Extended Data Fig. 6 | Survey error by age group (18–64-year-olds, and those aged 65 and over). a, Estimates of vaccine uptake from Delphi–Facebook (blue) and Census Household Pulse (green) for 18–64-year-olds (left) and those aged 65 or older (right). Bounds on the CDC's estimate of vaccine uptake for those groups are shown in grey. The CDC receives vaccination-by-age data only from some jurisdictions. We do know, however, the total number of vaccinations in the US. Therefore, we calculate the bounds by allocating all the vaccine doses for which age is unknown to either 18–64 or 65+. **b**, Unweighted ddc for each Delphi–Facebook and Census Household Pulse calculated for the 18–64 group using the bounds on the CDC's estimates of uptake. ddc for 65+ is not shown due to large uncertainty in the bounded CDC estimates of uptake.

Extended Data Table 1 | Methodologies of Axios–Ipsos, Census Household Pulse and Delphi–Facebook studies

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Purpose	Measure national attitudes toward COVID-19	Sub-national social and economic impact of COVID-19	Fine-grained COVID-19 symptom surveillance
Target Pop.	18+ US general pop	18+ US general pop	18+ US general pop
Length of wave	4 days, conducted weekly	2 weeks	Daily cross-section samples, reported weekly
Average participation rate among invitees	50%	6-8%	1%
Sampling design	Inverse response propensity sampling	Systematic sample of households, adjusted for a projected response rates	Unequal-probability stratified random samples
Hesitancy / Willingness question	"How likely, if at all, are you to get the first generation COVID-19 vaccine, as soon as it's available"	"Once a vaccine preventing COVID-19 is available to you, would you..."	"If a vaccine to prevent COVID-19 were offered to you today, would you choose to get vaccinated?"
Vaccine hesitancy responses	"Not very / at all likely"	"Definitely/Probably NOT get a vaccine" or "Unsure"	"No, definitely/probably not"
Languages	English and Spanish	English and Spanish	English, Spanish, Brazilian Portuguese, Vietnamese, French, and Chinese
Report MoE or design effect	Both	Report standard errors for estimates from replicate weights	Report standard errors for estimates (does not include variance from weighting)
Sources for demographic benchmarks	2019 CPS March Supplement, party ID from recent ABC/WaPo polls	2018 ACS, 1-year estimates	2018 CPS March Supplement

Extended Data Table 2 | Contribution of offline recruitment and weighting schemes to discrepancies between surveys

	Vaccinated		Hesitant	
	Raw	Weighted	Weighted	Sample size
Axios-Ipsos Survey				
only Offline Panelists	19%	13%	64%	21
only Online Panelists	43	37	30	974
with Ipsos Weights	42	36	30	995
with Delphi-implied Weights	42	37	29	995
Delphi-Facebook Survey				
with Delphi Weights	42%	46%	37%	249,954

A portion of each Axios–Ipsos wave is recruited from a population with no stable internet connection; Ipsos KnowledgePanel provides tablets to these respondents. In the Axios–Ipsos 22 March 2020 wave, the offline panellists (n = 21) were 24 percentage points less likely to be vaccinated than online panellists (n = 974). Weighting the same Axios–Ipsos data (n = 995) to the age and gender target distribution implied by Delphi–Facebook's weights make the vaccination estimates higher by 1 percentage point. However, this number is still lower than Delphi–Facebook's (responses from 14–20 March 2020, n = 249,954) own estimate of 46%. During this time period, the CDC benchmark vaccination rate was 35.2%. This suggests that the recruitment of offline respondents and different weighting schemes each explains only a small portion of the discrepancy between the two data sources.

Extended Data Table 3 | Example of multi-stage population selection

Stage	Population N_s	Sampling Process ⟶	Data n_s	$f_s = n_s/N_s$
1. Define frame	144 m hh	Subset to reachable address	116 m hh	80%
2. Decide outreach list	116 m hh	Random sample	1 m adults	1%
3. Individual behavior	1 m adults	Individual responds (or doesn't)	75,000	7%
Final	~ 255 m adults		75,000 adults	0.03%

The law of large populations described in the Methods section 'Population size in multi-stage sampling' shows that the population size at the sampling stage at which simple random sampling breaks down will dominate the error. This table explains these stages with a concrete example, using the Census Household Pulse. Population and sample sizes for three stages (stage number denoted s = 1, 2 or 3) of sampling of the Census Household Pulse survey data collection process. Approximate sample sizes based on the 24 March 2021 wave. 'm' stands for millions and 'hh' stands for household. The final row compares the total adult population in the US (255 million adults, made up of 144 million households) to the sample size in one wave of the household pulse. For illustration, we have ignored the effect of unequal sampling probabilities on the sample sizes at each stage.

nature portfolio

Corresponding author(s): Seth Flaxman

Last updated by author(s): Oct 4, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We collected the raw data from publicly available sources. No specific software was applicable.
Data analysis	Our code to download the public data and analyze it are all deposited at https://github.com/vcbradley/ddc-vaccine-US. All analysis was conducted in R (>= 3.6.0). All functions used are available at the Comprehensive R Archive Network (CRAN).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We analyze publicly available data from three surveys (available as microdata or summary statistics on their respective websites). An archive of the data we analyze is deposited in the Harvard Dataverse: https://doi.org/10.7910/DVN/GKBUUK. Delphi-Facebook provides summary statistics at https://www.cmu.edu/delphi-web/surveys/weekly-rollup/. Census Household Pulse provides microdata at https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html. Axios-Ipsos provides microdata at https://covid-19.parc.us.com/client/index.html#/search. The CDC benchmark is available at https://covid.cdc.gov/covid-data-tracker. The offline indicator of the Axios-Ipsos survey was provided to us and is included with permission in our Harvard Dataverse deposit.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The data are survey responses sampled by external organizations. It is quantitative in nature.
Research sample	We analyzed surveys by Delphi-Facebook, the Census Household Pulse, and Axios-Ipsos, and provide additional results comparing Data for Progress, Harris Poll, Morning Consult, and YouGov. We also analyzed administrative data on COVID-19 vaccinations reported by the CDC.
Sampling strategy	The three main surveys we analyzed were picked due to their frequency, large sample size, availability of summary statistics or microdata, large sample size, prominence in journalistic coverage and academic research, and the mode (online). The sampling strategies of each particular survey organization took is the main focus of the article and described in Table 1.
Data collection	We collected the survey data using download links and APIs from external, publicly available sources as described in the Data Availability Statement. The data collection strategies each particular survey organization took is the main focus of the article and described in Table 1.
Timing	We analyzed the survey waves taken from January 2021 to mid-May 2021.
Data exclusions	No data exclusions were made.
Non-participation	We take the survey estimates provided by the external providers as is and analyze its representativeness. The nature of the non-participation is a main focus of our methodological article.
Randomization	The surveys we analyzed are based in part on random samples; shortcomings are discussed in our methodological article.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging