Unrepresentative Big Surveys Significantly Overestimate US Vaccine Uptake

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

[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
[*]These authors contributed equally to this work.
[†]Correspondence to: seth.flaxman@cs.ox.ac.uk.

October 2021

Abstract

Surveys are a crucial tool for understanding public opinion and behavior, 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 impact 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: Delphi-Facebook[2,3] (about 250,000 responses per week) and Census Household Pulse[4] (about 75,000 per week). By May 2021, Delphi-Facebook overestimated uptake by 17 percentage points and Census Household Pulse by 14, compared to a benchmark from the Centers for Disease Control and Prevention (CDC). Moreover, their large data sizes led to minuscule margins of error on the incorrect estimates. In contrast, an Axios-Ipsos online panel[5] with about 1,000 responses following survey research best practices[6] provided reliable estimates and uncertainty. We decompose observed error using a recent analytic framework[1] to explain the inaccuracy in the three surveys. We then analyze 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 far more than data quantity, and 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 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 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] ($n \approx 250{,}000$ per week and with over 4.5 million responses from January to May 2021) and the Census Bureau's Household Pulse survey[4] ($n \approx 75{,}000$ 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, yet report significantly different estimates of vaccination behavior 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). In contrast, the Axios-Ipsos' Coronavirus Tracker[5] ($n \approx 1{,}000$

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].

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 the opportunity to examine the Big Data Paradox[1] in surveys. The Census Household Pulse is an experimental product designed to rapidly measure pandemic-related behavior. 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). Despite being the smallest survey by an order of magnitude, Axios-Ipsos' estimates track well with the CDC rates (see 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. Unfortunately, errors have increased over time, from just a few percentage points in January 2021 to 4.2 percentage points (Axios-Ipsos), 14 percentage points (Census Household Pulse), and 17 percentage points (Delphi-Facebook) by mid-May 2021 (Fig. 1b). For context, for a state 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. Fig. 1a shows 95% confidence intervals for vaccine uptake based on each survey's reported sampling standard errors and weighting design effects[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. Census Household Pulse's 95% confidence intervals 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 vanishingly small, driven by large sample size and moderate design effects (1.4-1.5), and give us essentially zero chance of being even close to the truth.

One benefit of such large surveys might be to compare estimates of spatial and demographic subgroups[24–26]. However, in March of 2021, Delphi-Facebook and Census Household Pulse over-estimated CDC state-level vaccine uptake by 16 and 9 percentage points, respectively (Extended Data Fig. 1G-H), 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 barely any 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, 0.26 for Census Household Pulse in Extended Data Fig. 1J) but they improved in May of 2021 (correlations 0.78 and 0.74 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 centers and substantially underestimated widths. This is the Big Data Paradox[1]: *the larger the data size, the surer we fool ourselves* when we fail to account for bias in data collection.

A framework for quantifying data quality

While it is well-understood that traditional confidence intervals capture only survey sampling errors[27] (and not total error), tools for quantifying nonsampling errors separately from sampling errors are difficult to apply and rarely used in practice. A recently formulated statistical framework[1] permits us to exactly decompose total error of a survey estimate into three components:

$$\textbf{Total Error} = \textbf{Data Quality Defect} \times \textbf{Data Scarcity} \times \textbf{Inherent Problem Difficulty}. \quad (1)$$

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, **Total Error** is the difference between the observed sample mean $\bar{Y}_n$ as an estimator of the ground truth, the population mean $\bar{Y}_N$. The **Data Quality Defect** is measured using $\hat{\rho}_{Y,R}$, called *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 impact of data quantity is captured by **Data Scarcity**, which is a function of the sample size n and the population size N, measured as $\sqrt{\frac{N-n}{n}}$, and hence what matters for error is the relative sample size, i.e., how close n is to N, rather the absolute sample size n. The third factor is **Inherent Problem Difficulty**, which measures the population heterogeneity (via 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 $\bar{Y}_n - \bar{Y}_N = \hat{\rho}_{Y,R} \times \sqrt{\frac{N-n}{n}} \times \sigma_Y$. Incidentally, this expression was inspired by the Hartley-Ross inequality for biases in ratio estimators published in *Nature* in 1954[30]. More details on the decomposition are provided in the Methods ("Calculation and interpretation of *ddc*"), where we also present a generalization for weighted estimators.

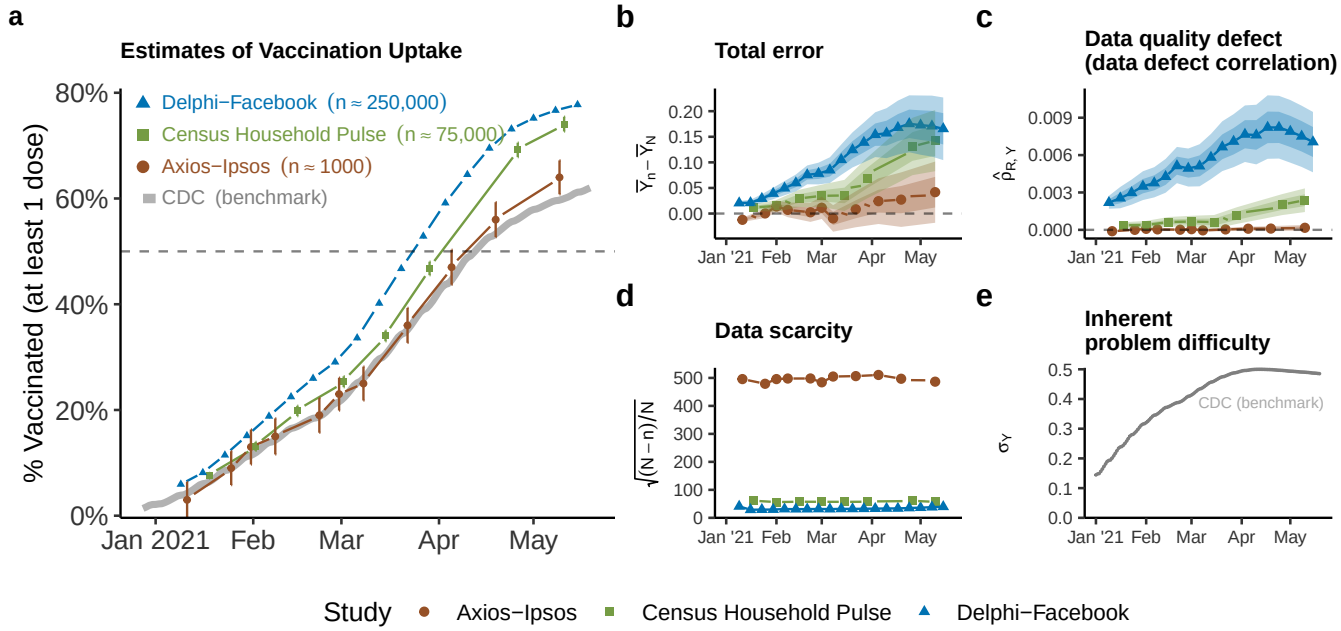


Fig. 1 | Errors in estimates of vaccine uptake. **a.** Estimates of vaccine uptake for US adults in 2021 compared to CDC benchmark data, plotted by end date of each survey wave. Points indicate each study's weighted estimate of first-dose vaccine uptake, and intervals are 95% CIs using reported standard 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 CIs 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) error in the CDC benchmark relative to reported values (points). **b** - **e** comprise the decomposition in Equation (1).

Decomposing error in COVID surveys

While the data defect correlation *ddc* is not directly observed, COVID-19 surveys present a rare case in which it can be deduced because all 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], $\bar{Y}_N$. This benchmark may suffer from administrative delays and slippage in how the CDC centralizes information from states[31–34]. As a sensitivity analysis to check the robustness of our findings to further misreporting, we present our results with sensitivity intervals under the assumption that CDC's reported numbers suffer from ±5% and ±10% error. These scenarios were chosen based on 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 and Supplementary Information A.2). That said, these scenarios may not capture latent systemic issues affecting CDC vaccination reporting.

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** ($\sqrt{\frac{N-n}{n}}$) for each survey is roughly constant across time (Fig. 1d). **Inherent Problem Difficulty** is a population quantity common to all three surveys which 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 impact 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 Square 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 impact of bias as well as that of an increase in variance from weighting and sampling. Details for this calculation are in Methods

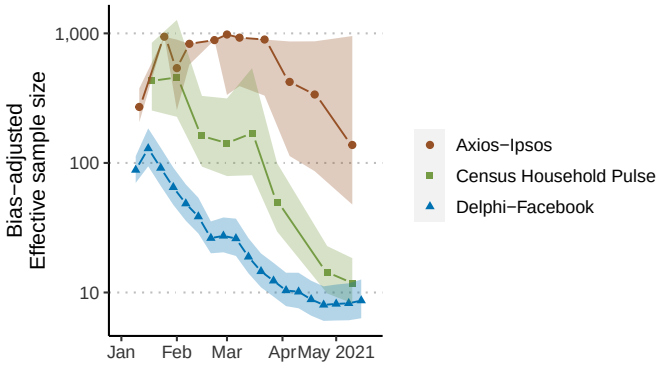


Fig. 2 | Bias-adjusted effective sample size. The bias-adjusted effective sample size of an estimate (different from the classic Kish effective sample size) is the size of a simple random sample which would have the same Mean Square Error as the observed estimate. Effective sample sizes are shown on 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, n = 11,421 across 11 waves for Axios-Ipsos. Shaded bands represent scenarios of +/-5% error in the CDC benchmark relative to point estimates based on actual reported values.

(Error decomposition with survey weights).

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 also suffers from 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 drastic reductions of effective sample sizes for large populations[1].

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 the "Additional survey methodology" in

	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 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

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

Methods and Extended Data Table 1).

All three surveys are conducted online and target the US adult population, but vary in respondent recruitment methods[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 Census' Master Address File (MAF) for which they have obtained either cell phone or email contact information (approximately 81% of all households on the MAF)[4].

In comparison, Axios-Ipsos relies on inverse response propensity sampling from Ipsos'

online KnowledgePanel. Ipsos recruits panelists using an address-based probabilistic sample from USPS's Delivery Sequence File (DSF)[5]. The DSF is similar to the Census' MAF. 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 panelists who lack home internet access. In 2021, this "offline" group typically comprises 1% of the final survey (Extended Data Table 2).

All three surveys weight on age and gender, i.e., 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/ethnicity. Axios-Ipsos additionally weights to the composition of political partisanship measured with the ABC News/Washington Post poll 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, while 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 story is similar for race/ethnicity. Delphi-Facebook's weighting scheme does not adjust for race/ethnicity, and hence their weighted sample still over-represents White adults by 8 percentage points, and under-represents Black and Asian proportions by around 50 percent of their size in

	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
4-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

Table 2 | Composition of survey respondents by educational attainment and race/ethnicity. Axios-Ipsos: wave ending March 22, 2021, n = 995. Census Household Pulse: wave ending March 29, 2021, n = 76,068. Delphi-Facebook: wave ending March 27, 2021, n = 181,949. Benchmark uses the 2019 US Census American Community Survey (ACS), composed of roughly 3 million responses. Right-most column shows estimates of vaccine uptake (Vax), willingness (Will) and hesitancy (Hes) from the Census Household Pulse of the same wave.

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 under-represents tend to be more willing and less vaccinated in the samples (Table 2). In other words, re-weighting 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 4-year college degree are less likely to have been vaccinated compared to those *with* a degree (Table 2 and Supplemental Information Table S1). If we assume that vaccination behaviors do not differ systematically between non-respondents and respondents *within* each demographic category, under-representation 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 over-estimates 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 behavior[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; Census agencies are prohibited from asking about political preference. Moreover, no unequivocal population benchmark for partisanship exists.

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

Axios-Ipsos weights to metropolitan status and also recruits a fraction of its panelists from an "offline" population of individuals without Internet access. We find that *dropping* these offline respondents (n = 21, or 1 percent of the sample) in their March 22, 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 simply too small to explain the entirety of the difference in accuracy between Axios-Ipsos and either the Census Household Pulse (6 percentage points) or Delphi-Facebook (14 percentage points), in this time period.

Careful recruitment of panelists 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 March 22, 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 weight-

ing 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 instance the selection of participants followed by non-response, bias in even a *single* step can substantially impact the final result (see Methods "Population size in multi-stage sampling", 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 in order to improve data quality. This emphasis is merely 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 far more statistically confident, yet also far more biased, than the smaller, more traditional Axios-Ipsos poll. These findings are paradoxical only when we fall into the trap of the long-held, but incorrect, 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 expose 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 via probabilistic sampling[43]. Indeed, Neyman's original introduction of probabilistic sampling showed the limits of purposive sampling, which attempted to achieve overall representativeness via enforcing it only on a set of variables[18,44].

In other words, when a survey loses its overall statistical representativeness (e.g., through bias in coverage or nonresponse), which is difficult to repair (e.g., via weighting or modeling 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 wish to publish studies based on unrepresentative surveys[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. A simple acknowledgment of the potential bias is insufficient for alerting about potentially seriously flawed findings, as we reveal in this article.

Some may argue that bias is a necessary trade-off for having data that is sufficiently large for conducting highly granular analysis, such as county-level estimation of vaccine hesitancy[26]. While 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 also relies 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 yet another example of the Big Data Paradox.

Discussion

This is not the first time that the Big Data Paradox has reared its head: 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, yet overestimate vaccine uptake by double digits. As we demonstrated, the impact of bias is magnified as relative sample size increases.

In 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 analyzed, 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 (i.e., leading to unhelpfully large confidence intervals) for the kind of 50-state estimates given by big surveys. However, small area methods that borrow information across subgroups[48] can perform better with better quality, albeit small, data, and it is an open question whether that approach would outperform the large, biased surveys.

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 multi-mode 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 (Supplemental Information D).

While more needs to be done to fully examine the nuances of large surveys, organically collected administrative datasets, and social media data, we hope this first comparative study of *ddc* highlights the alarming implications of the *Big Data Paradox* – how large sample sizes magnify the impact of seemingly small defects in data collection, leading to overconfidence in incorrect inferences.

Methods

Calculation and interpretation of *ddc*

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

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

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 $\sigma_Y = \sqrt{\bar{Y}_N\left(1-\bar{Y}_N\right)}$ 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 (e.g., 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 Methods section "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 behavior rather than eligibility, the correlation between

vaccination status and propensity to respond could increase even if 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 $\bar{Y}_w = \sum_i w_i R_i Y_i / \sum_i w_i R_i$, where w_i is the survey weight assigned to individual i. It is shown in Meng[1] that

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

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], i.e., $n_{\mathrm{w}} = n/(1+CV_{\mathrm{w}}^2)$, 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 impact 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 data set 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 to reduce the impact 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 $f = n/N$ to $f_w = n_w/N$ as well since

$n_w < n$. 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 via a probabilistic scheme, or weighted to mimic a probabilistic sample. As discussed in the the section on addressing common missperception, any single *ddc* deviating from what is expected under representative sampling (e.g., probabilistic sampling) is sufficient to establish 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}:

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

Given an estimator $\bar{Y}_{\text{w}}$ with expected total Mean Squared Error (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 $E[\hat{\rho}^2_{Y,R_{\text{w}}}]$ represents the amount of selection bias (square) expected on average from a particular recording mechanism R and a chosen weighting scheme.

For each survey wave, we use $\hat{\rho}^2_{Y,R_{\text{w}}}$ to approximate $E[\hat{\rho}^2_{Y,R_{\text{w}}}]$. This estimation is unbiased

by design, since 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 for Delphi-Facebook and the Census Household Pulse surveys. This estimation error may have more impact for smaller traditional surveys, such as Axios-Ipsos' survey, an issue we will investigate in subsequent work.

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

$$\hat{\rho}_{Y,R_{\mathrm{w}}} = \frac{Z_w}{\sqrt{N}}, \quad \text{where} \quad Z_w = \frac{\bar{Y}_{\mathrm{w}} - \bar{Y}_N}{\sqrt{\frac{1-f_{\mathrm{w}}}{n_{\mathrm{w}}}}\sigma_Y}. \tag{4}$$

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 where $\bar{Y}_w$ may be biased, we replace the variance by MSE, and hence the bias-adjusted design effect $D_e = E[Z_w^2]$, 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 $D_I = D_e/N$.

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, $Z_w/\sqrt{N}$ no longer has the simple interpretation as a correlation. But because we estimate D_I by Z_w^2/N directly, our effective sample size calculation is still valid even when Equation (3) does not hold.

Asymptotic behavior of *ddc*

As shown in Meng[1], for any probabilistic sample without selection biases, the *ddc* is on the order of $1/\sqrt{N}$. Hence the magnitude of $\hat{\rho}_{Y,R}$ (or $\hat{\rho}_{Y,R_{\mathrm{w}}}$) is small enough to cancel out the impact 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, e.g. 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 $1/\sqrt{N}$ 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 counter-intuitive in the traditional survey statistics framework, which often considers how error changes as sample size n grows. The *ddc* framework considers a more general setup, taking into account individual response behavior, including its impact on sample size itself.

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

$$\text{logit}[\Pr(R = 1|Y)] = \alpha + \beta Y. \tag{5}$$

This is a model for a response propensity score. Its value is determined by α, which drives the overall sampling fraction $f = n/N$, 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 behavior, not by population size N. In Supplemental 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 behavior 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, while the MSE for the latter (a -4.84 coefficient on Y) is 0.242, over 600 times larger.

See Supplemental 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 behavior of error depends on whether the data collection process is driven by individual response behavior 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, i.e., when the individual response behaviors come into play. Real-world surveys that are as complex as the three surveys we analyze 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 analyze. As we have summarized (Table 1 and Extended Data Table 1), the Census Household Pulse (1) first defines the sampling frame as the reachable subset of the Master Address File, (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, for stage 3, the population size N_3 is the size of the intended sample size n_2 from the second stage, i.e., the sampling stage, because only the sampled individuals have a chance to respond.

Although all stages contribute to the grand *ddc*, the stage that dominates is the *first stage at which the representativeness of our sample is destroyed*— whose size will be labeled as the *dominating population size (dps)*—when the relevant population size decreases dramatically at each step. However, we must bear in mind that *dps* refers to the worse case scenario, when biases accumulate, instead of (accidentally) cancel each other out.

For example, if the 20 percent of the MAF excluded from the Census Household Pulse sampling frame (because they had no cell phone or email contact information) is not represen-

tative 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 $\sqrt{2.5 \times 10^8} \approx 15,000$). In 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 *ddc* is amplified by the square root of that number ($\sqrt{10^6} = 1,000$). In contrast, Axios-Ipsos reports a response rate of about 50%, and obtains a sample of $n = 1000$, so the *dps* could be as small as $N_3 = 2000$ (with $\sqrt{2000} \approx 45$).

This decomposition is why our comparison of the surveys is consistent with the *Law of Large Populations* (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. In contrast, the main source of unrepresentativeness for Axios-Ipsos maybe at a later stage where the relevant population size is orders of magnitude smaller.

CDC estimates of vaccination rates

The CDC benchmark data used in our analysis was downloaded from the CDC's COVID data tracker[13]. We employ the cumulative count of people who have received at least one dose of COVID-19 vaccine reported in the "Vaccination Trends" tab. This data set contains vaccine uptake counts for all US residents (not only adults). However, the surveys of interest estimate vaccine uptake among adults. The CDC receives age-group-specific data on vaccine uptake from all states except for Texas on a daily basis, which is also reported cumulatively over time.

Therefore, we must impute the number of adults who have received at least one dose on each

day. We assume Texas is exchangeable with the rest of the states in terms of the age distribution for vaccine uptake. Under this assumption, for each day, we use the age group vaccine uptake data from all states except for Texas to calculate the proportion of cumulative vaccine recipients who are 18 or older, then we multiply that number by the total number of people who have had at least one dose to estimate the number of US *adults* who have received at least one dose.

The CDC performs a similar imputation for the 18+ numbers reported in their COVID data tracker. However the CDC's imputed 18+ number is available only as a snapshot and not a historical time series, hence the need for our imputation. See Supplemental Information for details of the imputation implementation.

Additional survey methodology

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

Question wording All three surveys ask whether respondents have received a COVID-19 vaccine. See 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 behavior and thus misreport their true vaccination status, or may induce acquiescence bias from the multiple "yes" options presented. This pressure may exist both in high- and low-vaccination communities, so its net impact on Axios-Ipsos' results is unclear. Nonetheless, Axios-Ipsos' question wording does differ from that of the other two surveys, and may contribute the observed differences in estimates of vaccine uptake across surveys.

Population of Interest All three surveys target US adult population, but with different sampling and weighting schemes. 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 representative of the US general adult population 18 or older. For 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 March 13, 2020 and has released results from 45 waves as of May 28, 2021. Each wave generally runs over a period of 4 days. The Axios-Ipsos data used in this analysis was 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 $1 + CV_{\mathrm{w}}^2$, 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 a 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 was downloaded directly from CMU's repository for weekly contingency tables with point estimates and standard errors.

References

[1] Xiao-Li Meng. 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, 2018.

[2] Neta Barkay, Curtiss Cobb, Roee Eilat, Tal Galili, Daniel Haimovich, Sarah Larocca, Katherine Morris, and Tal Sarig. Weights and methodology brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in partnership with Facebook. https://arxiv.org/abs/2009.14675, 2020.

[3] Frauke Kreuter *et al.* Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods*, 14(2):159–163, 2020. doi: 10.18148/srm/2020.v14i2.7761.

[4] Jason F. Fields, Jeniffer Hunter-Childs, A. Tersine, J. Sisson, E. Parker, V. Velkoff, C. Logan, and H. Shin. Design and operation of the 2020 Household Pulse survey. 2020. U.S. Census Bureau. https://perma.cc/JC3D-3LBY.

[5] Chris Jackson, Mallory Newall, and Jinhee Yi. Axios Ipsos Coronavirus Index, 2021. https://www.ipsos.com/en-us/news-polls/axios-ipsos-coronavirus-index.

[6] American Association for Public Opinion Research (AAPOR). Best practices for survey research. 2021. URL https://perma.cc/3M56-668Q.

[7] Manoj Hastak, Michael B Mazis, and Louis A Morris. The role of consumer surveys in public policy decision making. *Journal of Public Policy & Marketing*, 20(2):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. *Morbidity and Mortality Weekly Report*, 2021. doi: 10.15585/mmwr.mm7020e3. https://perma.cc/FV7A-USGC.

[9] Alejandra Arrieta, Emmanuela Gakidou, Heidi Larson, Erin Mullany, and Christopher Troeger. Through understanding and empathy, we can convince women to get the covid-19 vaccine. *Think Global Health*, 2021.

[10] Lilli Japec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. AAPOR report on Big Data. *American Association of Public Opinion Researchers*, 2015. URL https://perma.cc/2X5U-LNN7.

[11] Alex Reinhart, Esther Kim, Andy Garcia, and Sarah LaRocca. Using the COVID-19 Symptom Survey to track vaccination uptake and sentiment in the United States, 2021. Sept 30, 2021 https://perma.cc/GB72-C6Q5.

[12] Viktor Mayer-Schönberger and Kenneth Cukier. *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. US Centers for Disease Control (CDC), https://covid.cdc.gov/covid-data-tracker/#vaccination-trends.

[14] Kimberly H. Nguyen, Peng-Jun Lu, Seth Meador, Mei-Chuan Hung, Katherine Kahn, Jessica Hoehner, Hilda Razzaghi, Carla Black, and James A. Singleton. 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. https://www.cdc.gov/vaccines/imz-managers/coverage/adultvaxview/pubs-resources/covid19-coverage-estimates-comparison.html.

[15] Tammy A. Santibanez, James A. Singleton, Carla L. Black, Kimberly Nguyen, Mei-Chuan Hung, Svetlana Masalovich, Peng-Jun Lu, Kathryn A. Brookmeyer, Neetu Abad, Kamil E. Barbour, Ari Whiteman, Bhavini Patel Murthy, Alice Wang, and Holly A. Hill. 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. 2021. https://perma.cc/C6KN-UDUY.

[16] William Kruskal and Frederick Mosteller. Representative sampling, I: Non-scientific literature. *International Statistical Review/Revue Internationale de Statistique*, pages 13–24, 1979.

[17] William Kruskal and Frederick Mosteller. Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review/Revue Internationale de Statistique*, pages 111–127, 1979.

[18] William Kruskal and Frederick Mosteller. Representative sampling, III: The current statistical literature.

[19] William Kruskal and Frederick Mosteller. Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, pages 169–195, 1980.

[20] AAPOR. Margin of Sampling Error / Credibility Interval. Oct 23, 2021 https://perma.cc/Q79L-C28F.

[21] The Delphi Group at Carnegie Mellon University in partnership with Facebook. Topline Report on COVID-19 Vaccination in the United States, 2021. Sept 30, 2021 https://perma.cc/FKU8-LSYD.

[22] Eric J. Haas, Frederick J. Angulo, John M. McLaughlin, Emilia Anis, Shepherd R. Singer, Farid Khan, Nati Brooks, Meir Smaja, Gabriel Mircus, Kaijie Pan, 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. *The Lancet*, 2021.

[23] Leslie Kish. *Survey Sampling*. Wiley, 1965. ISBN 0-471-10949-5.

[24] Institute for Health Metrics and Evaluation (IHME). Covid-19 vaccine hesitancy, 2021. Oct 3, 2021 https://vaccine-hesitancy.healthdata.org.

[25] Wendy C. King, Max Rubinstein, Alex Reinhart, and Robin J. Mejia. Time trends and factors related to COVID-19 vaccine hesitancy from January-May 2021 among US adults: Findings from a large-scale national survey. *medRxiv*, 2021. https://doi.org/10.1101/2021.07.20.21260795.

[26] Centers for Disease Control. Estimates of vaccine hesitancy for COVID-19, 2021. Sept 30, 2021 https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw/.

[27] Robert M. Groves, Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*, volume 561. John Wiley & Sons, 2011.

[28] Walter Dempsey. The hypothesis of testing: Paradoxes arising out of reported Coronavirus case-counts. https://arxiv.org/abs/2005.10425, 2020.

[29] Michael Isakov and Shiro Kuriwaki. Towards principled unskewing: Viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review*, 2(4), 2020.

[30] Herman Otto Hartley and Arun Ross. Unbiased ratio estimators. *Nature*, 174(4423): 270–271, 1954.

[31] Andrew Tiu, Zachary Susswein, Alexes Merritt, and Shweta Bansal. Characterizing the spatiotemporal heterogeneity of the COVID-19 vaccination landscape. 2021. Preprint at http://bansallab.com/COVID_County_vaccination.pdf.

[32] Jeffery Groen. Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics*, 28:173–198, 2012.

[33] Xin Ming Tu, Xiao-Li Meng, and Marcello Pagano. The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association*, 88(421):26–36, 1993.

[34] Oliver Barnes and John Burn-Murdoch. COVID response hampered by population data glitches. *Financial Times*, 2021. Oct 13, 2021 https://www.ft.com/content/125fbaf8-175a-4e2e-852a-9995ca5176b2.

[35] Courtney Kennedy, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley McGeeney, and Alejandra Gimenez. Evaluating online nonprobability surveys. *Pew Research Center*, 2016.

[36] Courtney Kennedy, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, and Douglas Rivers. An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1):1–33, 2018. doi: 10.1093/poq/nfx047.

[37] Brooke Auxier and Monica Anderson. Social media use in 2021. *Pew Research Center*, 2021.

[38] Shana K. Gadarian, Sara W. Goodman, and Thomas B. Pepinsky. Partisanship, health behavior, and policy attitudes in the early stages of the COVID-19 pandemic. *PLOS One*, 16(4), 2021.

[39] Andrew Mercer, Arnold Lau, and Courtney Kennedy. For Weighting Online Opt-In Samples, What Matters Most? *Pew Research Center*, 2018.

[40] Camille Ryan. Computer and internet use in the United States: 2016. *American Community Survey Reports*, ACS-39, U.S. Census Bureau, Washington, DC, 2017.

[41] Paul P. Biemer and Lars E. Lyberg. *Introduction to survey quality*. John Wiley & Sons, 2003.

[42] Fritz Scheuren. *What is a Survey?* American Statistical Association, 2004.

[43] Pandurang V Sukhatme. Sampling theory of surveys with applications. Technical report, 1954.

[44] Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of Royal Statistical Society*, 97:558–625, 1934.

[45] Robert M Groves. Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, 70(5):646–675, 2006.

[46] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014. ISSN 10959203. doi: 10.1126/science.1248506.

[47] Joshua A Salomon, Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichida La Motte-Kerr, Minttu Rönn, Marissa B Reitsma, Katherine Ann Morris, Sarah LaRocca, Tamer

Farag, Frauke Kreuter, Roni Rosenfeld, and Ryan J. Tibshirani. The US COVID-19 Trends and Impact Survey, 2020-2021: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing and vaccination. 2021. Preprint at https://doi.org/10.1101/2021.07.24.21261076.

[48] David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4): 375–385, 2004.

[49] Associated Press-NORC Center for Public Affairs Research. The june 2021 ap-norc center poll. 2021. July, https://perma.cc/6ZXM-58XT.

[50] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[51] Michael R Elliott and Richard Valliant. Inference for nonprobability samples. *Statistical Science*, 32(2):249–264, 2017.

[52] Roderick J.A. Little, Brady T. West, Philip S. Boonstra, and Jingwei Hu. Measures of the degree of departure from ignorable sample selection. *Journal of survey statistics and methodology*, 8(5):932–964, 2020.

[53] Arkadiusz Wiśniowski, Joseph W. Sakshaug, Diego Andres Perez Ruiz, and Annelies G Blom. Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1):120–147, 2020.

[54] Shu Yang, Jae Kwang Kim, and Rui Song. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):445–465, 2020.

[55] Methodology for the United States population estimates: Vintage 2019. US Census Bureau, https://perma.cc/PCC4-V48Q.

[56] Jelke Bethlehem. *Weighting Nonresponse Adjustments Based on Auxiliary Information*. New York: Wiley, 2002. In "Survey Nonresponse", ed. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little.

[57] Xiao-Li Meng. *A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it)*. CRC Press, 2014.

[58] Xiao-Li Meng and Xianchao Xie. I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews*, 33(1-4):218–250, 2014.

[59] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, pages 153–161, 1979.

[60] CDC. Reporting COVID-19 vaccination demographic data, 2021. US Centers for Disease Control (CDC), https://perma.cc/H8A5-D7RX.

[61] Cameron Appel, Diana Beltekian, Daniel Gavrilov, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Edouard Mathieu, Esteban Ortiz-Ospina, Hannah Ritchie, Lucas Rodés-Guirao, and Max Roser. Data on COVID-19 (coronavirus) by Our World in Data, 2021. https://github.com/owid/covid-19-data.

[62] Peter Bouman, Vanja Dukic, and Xiao-Li Meng. A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*, pages 325–357, 2005.

[63] Mark Blumenthal. Why YouGov is changing how we ask people whether they've received the COVID-19 vaccine. 2021. May 4, https://perma.cc/2EYN-K358.

[64] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16:101–117, 2001.

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.

Data availability

Raw data is deposited in the Harvard Dataverse, at https://doi.org/10.7910/DVN/GKBUUK. Data was collected from publicly available repositories of survey data by downloading it 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).

Acknowledgments

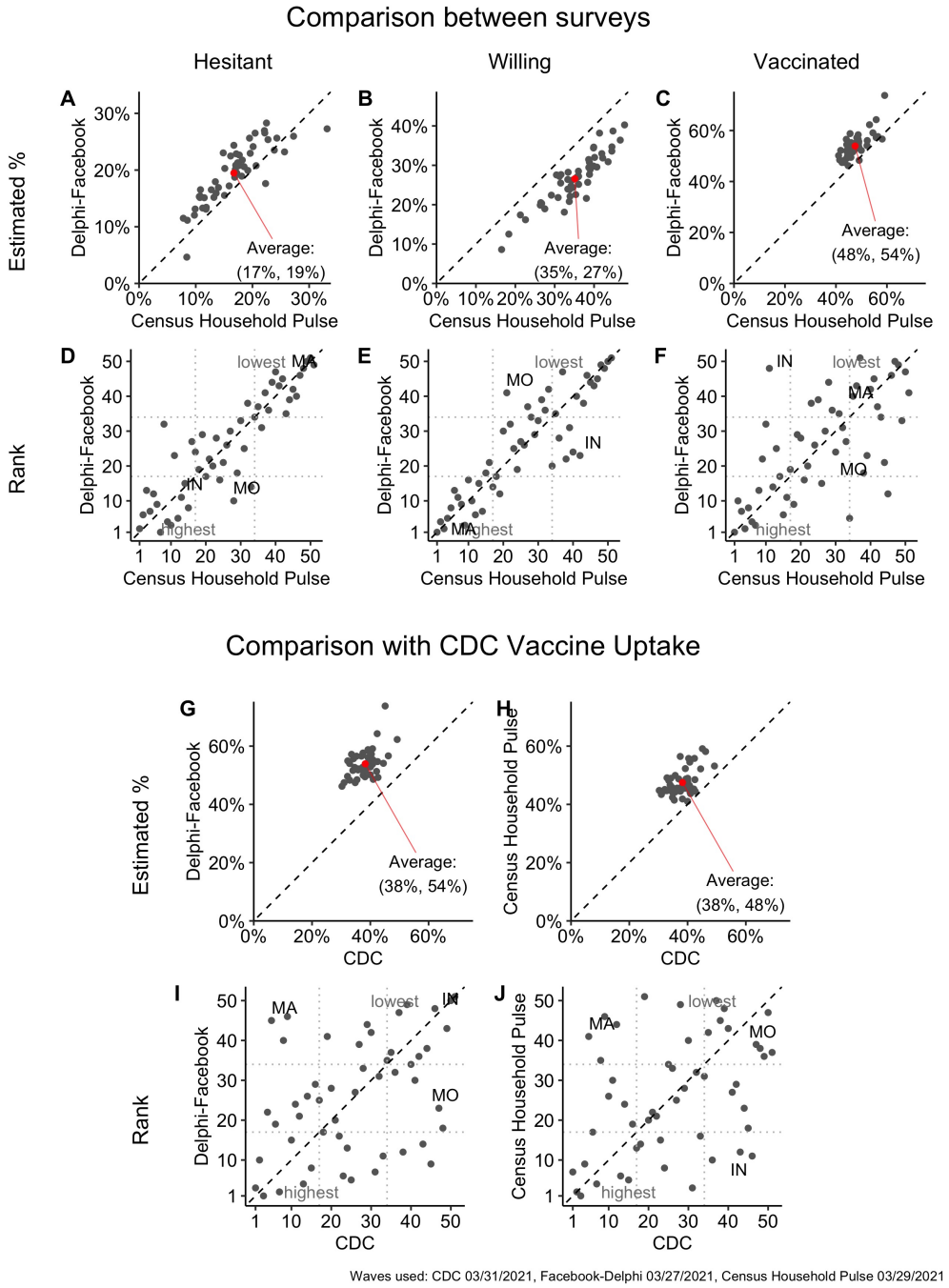
We thank Frauke Kreuter, Alex Reinhart, and the Delphi Group at Carnegie Mellon University, Facebook's Demography and Survey Science group; Frances Barlas, Chris Jackson, Catherine Morris, Mallory Newall, and the Public Affairs team at Ipsos; and Jason Fields and Jennifer Hunter Childs at the US Census 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, and the CDC for

responding to our questions. Susan Paddock, other participants at the JPSM 2021 lecture (delivered by Meng), and Steve Finch provided helpful comments, which we greatly appreciate. We thank the anonymous reviewers for their constructive comments, which substantially improved our work. We thank Ariel Edwards-Levy for a tweet which originally inspired our interest in this topic, and Rick Born for suggesting more intuitive terms used in Equation (1). V.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).

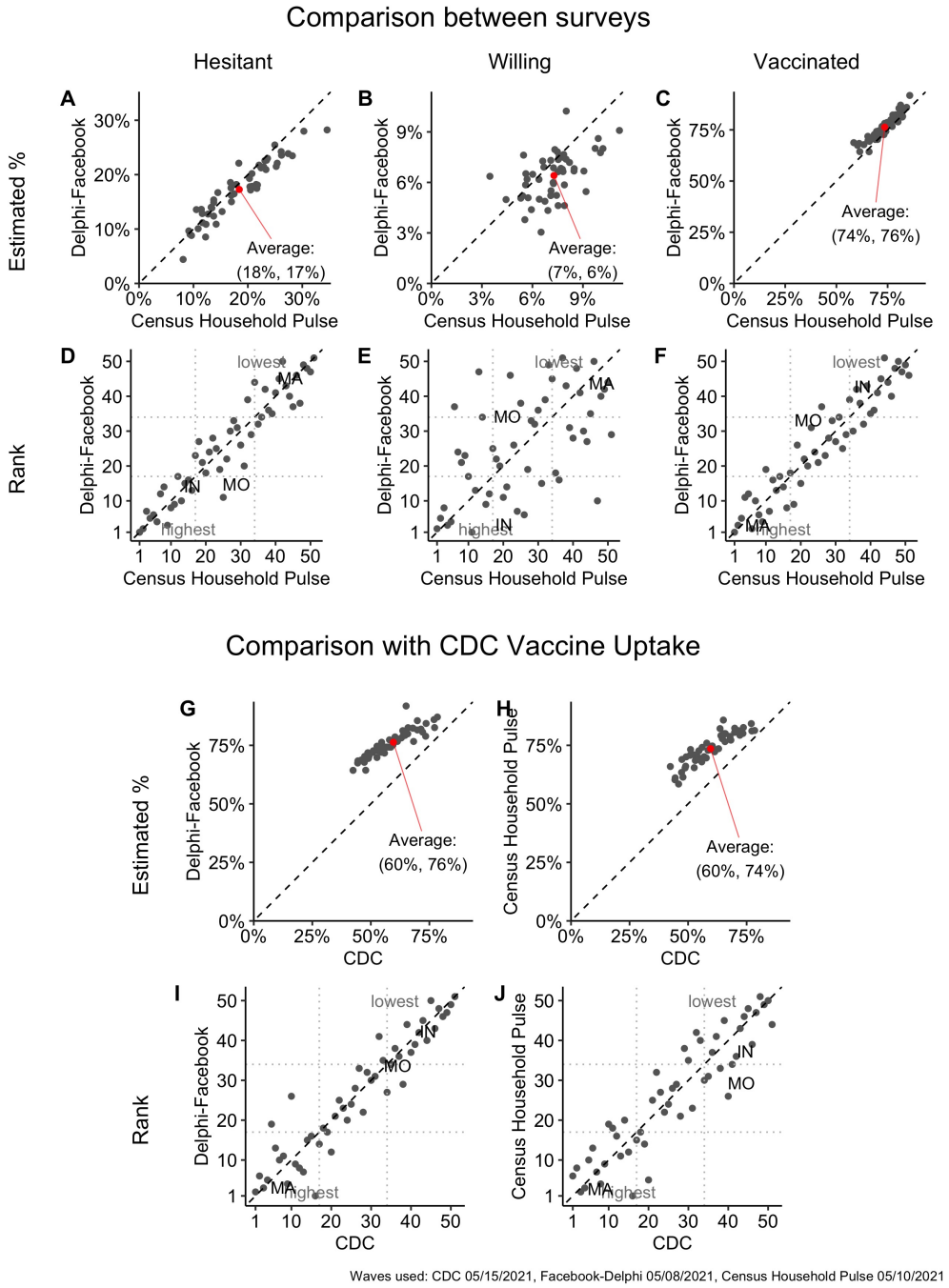
Author contributions V.B. and S.F. conceived and formulated the research questions. V.B. and S.K. contributed equally to data analysis, writing, and visualization. X-L.M. conceived and formulated the methodology. All authors contributed to methodology, writing, visualization, editing, and data analysis.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

Competing Interests Authors have no competing interests, financial or otherwise.

Extended Data



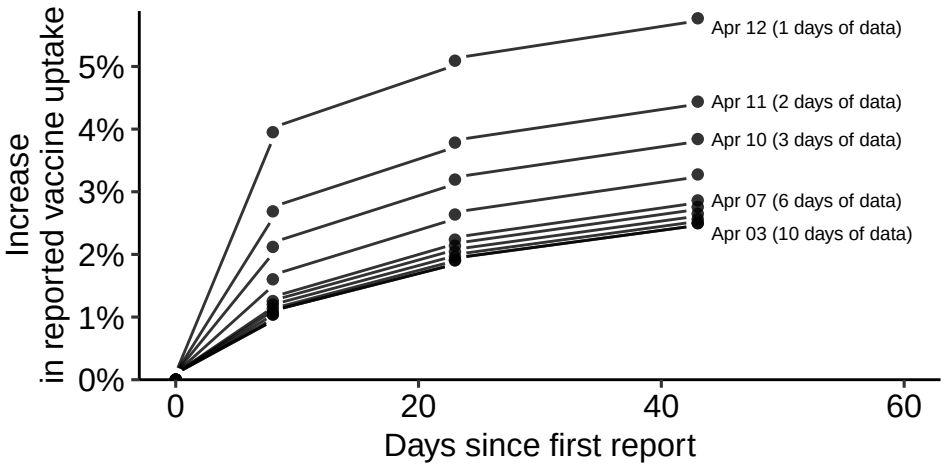
Extended Data Fig. 1 | Comparisons of state-level vaccine uptake, hesitancy, and willingness across surveys and the CDC: March 2021 Comparison of state-level point estimates (A-C) and rankings (D-F) for vaccine hesitancy, willingness, and uptake from Delphi-Facebook, and Census Household Pulse. Dotted black lines show agreement and red points show the average of 50 states. Panels G-J compare state-level point estimates and rankings for the same survey waves to CDC benchmark estimates from March 31, 2021. The Delphi-Facebook data is from the week ending March 27, 2021 and the Census Household Pulse is the wave ending March 29, 2021.



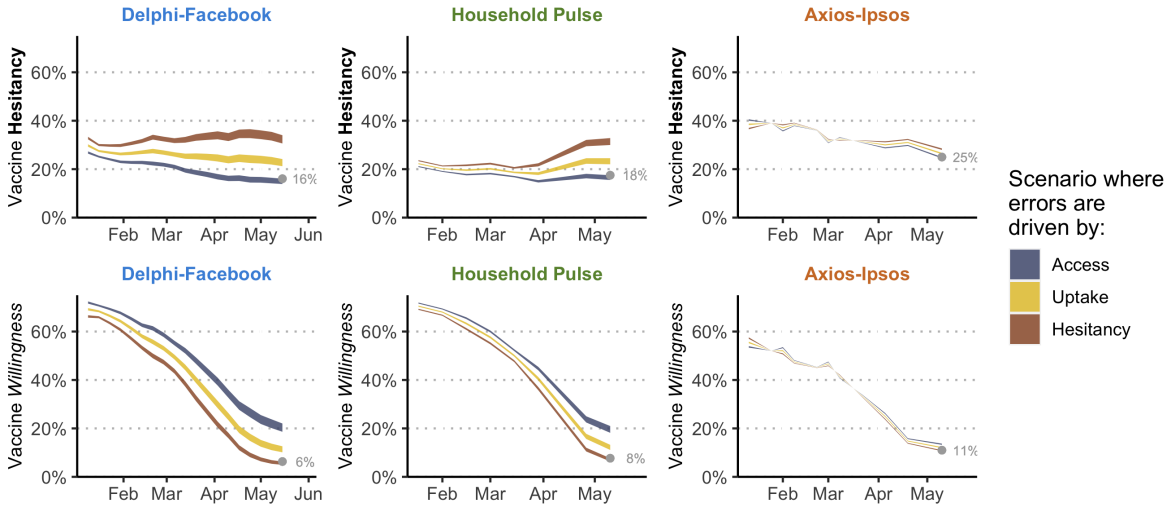
Extended Data Fig. 2 | Comparisons of state-level vaccine uptake, hesitancy, and willingness across surveys and the CDC: May 2021. Comparison of state-level point estimates (A-C) and rankings (D-F) for vaccine hesitancy, willingness, and uptake from Delphi-Facebook, and Census' Household Pulse. Dotted black lines show agreement and red points show the average of 50 states. Panels G-J compare state-level point estimates and rankings for the same survey waves to CDC benchmark estimates from May 15, 2021. The Delphi-Facebook data is from the week ending May 8, 2021 and the Census Household Pulse is the wave ending May 10, 2021.

	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

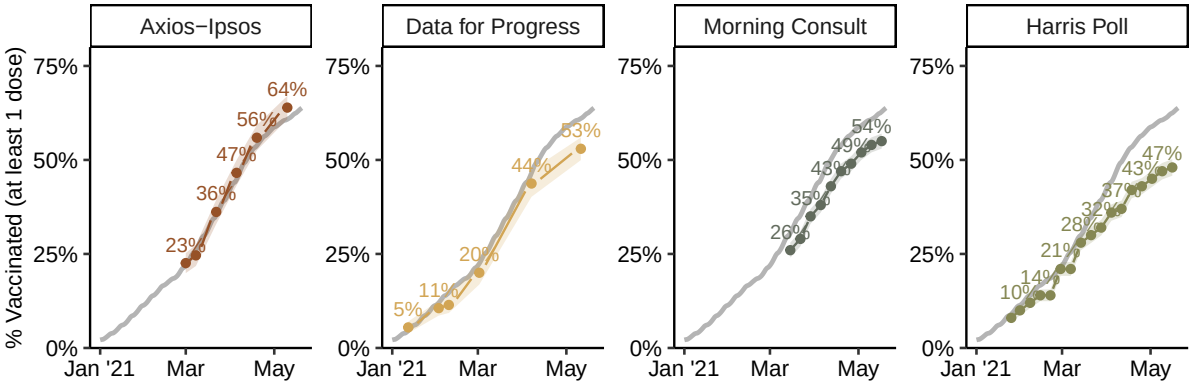
Extended Data Table 1 | Methodologies of Axios-Ipsos, Census Household Pulse, and Delphi-Facebook studies. Supplements information in Table 1.



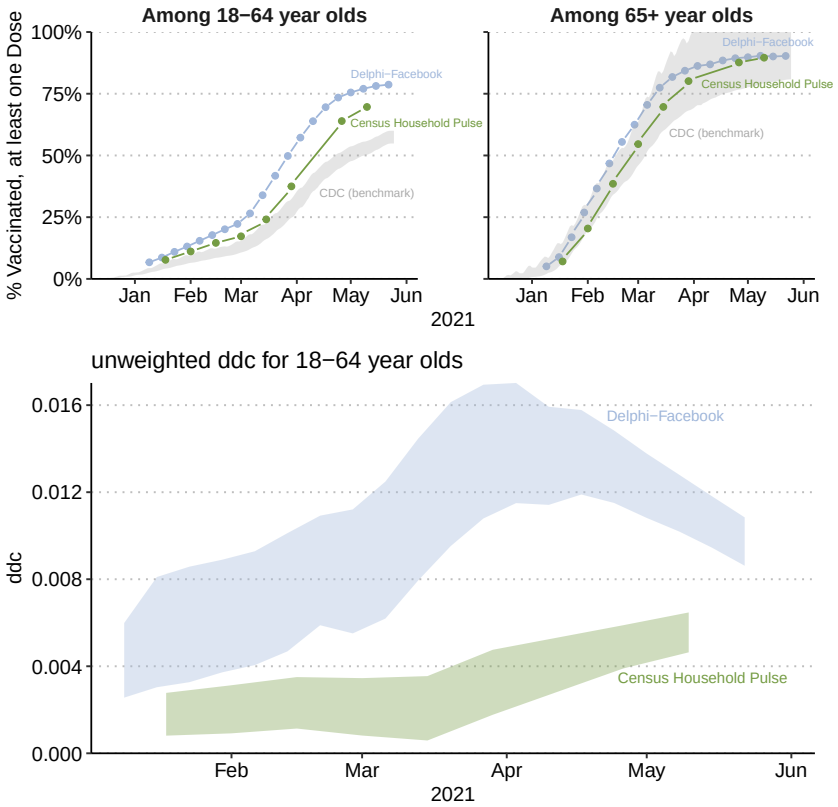
Extended Data Fig. 3 | Retroactive adjustment of CDC vaccine uptake figures for April 3-12, 2021, over the 90 days from April 12. Increase is shown as a percentage of the vaccine uptake reported on April 12. Most of the retroactive increases in reported estimates appear to occur in the first 10 days 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% error in the CDC benchmark for our robustness checks.



Extended Data Fig. 4 | Revised estimates of hesitancy and willingness after accounting for survey errors for vaccination uptake. The gray 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 percent confidence intervals which are twice the standard error reported by the poll. 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?". See the Supplementary Information C.3 for more details on each survey and discussion of differences. Gray line is the CDC benchmark.



Extended Data Fig. 6 | Survey error by Age Group (18-64 year-olds, and those 65 and over). **a.** Estimates of vaccine uptake from Delphi-Facebook (blue) and Census Household Pulse (green) for each 18-64 year-olds (left) and those 65 or older (right). Bounds on the CDC's estimate of vaccine uptake for those groups are shown in gray. The CDC receives vaccination-by-age data only from some jurisdictions. We do know, however, the total number of vaccinations in the U.S. Therefore, we calculate the bounds 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.

	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

Extended Data Table 2 | Contribution of offline recruitment and weighting schemes to discrepancies between surveys. 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 March 22 2020 wave, the offline panelists (n = 21) were 24 percentage points less likely to be vaccinated than online panelists (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 March 14–20 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.

Stage s	**Population N_s**	**Sampling Process $\longrightarrow$**	**Data n_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 adults	7%
Final	~ 255 m adults		75,000 adults	0.03%

Extended Data Table 3 | Example of multi-stage population selection. The *Law of Large Populations* described in Methods section "Population size in multi-stage sampling" shows that the population size at the sampling stage where 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 $s \in \{1, 2, 3\}$) of sampling of the Census Household Pulse survey data collection process. Approximate sample sizes based on the March 24, 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 the purpose of illustration, we have ignored the impact of unequal sampling probabilities on the sample sizes at each stage.

Supplementary Information

Contents

A Additional Details about Data Sources **2**
- A.1 Total Population 2
- A.2 CDC Imputation and Uncertainty 2
- A.3 Availability of Survey Microdata 4
- A.4 Census Household Pulse 4
- A.5 Delphi-Facebook 5

B Asymptotic Properties of *ddc* **5**
- B.1 The Role of Individual Response Behavior 6
- B.2 Connection with the Heckman selection model 7

C Additional Data Analyses **9**
- C.1 Estimates of hesitancy by demographic groups 9
- C.2 *ddc* by age / eligibility status across time 9
- C.3 Other online polls 10

D *ddc*-based Scenario Analysis for Willingness and Hesitancy **14**
- D.1 Setting up scenarios 14
- D.2 Obtaining scenario estimates 15
- D.3 Scenario estimates 16

A Additional Details about Data Sources

A.1 Total Population

The CDC vaccination data includes vaccines administered in Puerto Rico. As of June 9, 2021, approximately 1.6 million adults have received at least one dose, just under 1% of the national total (164,576,933). We use the CDC's reported national total that includes Puerto Rico (we do not have a reliable state-level time series of vaccine uptake), but we use a denominator that *does not* include Puerto Rico. This means that the CDC's estimate of vaccine uptake used here may be slightly *overestimating* the true proportion of the US (non-Puerto Rico) adult population that has received at least one dose by about 1%, which would make the observed *ddc* for Delphi-Facebook and Census Household Pulse and *underestimate* of the truth. However, this 1% error is well within the benchmark uncertainty scenarios presented with our results.

A.2 CDC Imputation and Uncertainty

The CDC does release state-level snapshots of vaccine uptake each day. These have been scraped and released publicly by Our World In Data[61]. These state-level numbers are not historically-updated as new reports of vaccines administered on previous days are reported to the CDC, so they underestimate the true rate of state-level vaccine uptake on any given day. These data are used only to motivate the inaccuracies of the state-level rank orders implied by vaccine uptake estimates from Delphi-Facebook and Census Household Pulse; hence they are not used to calculate *ddc*.

To inform our CDC benchmark uncertainty scenarios, we examined changes in vaccine uptake rates reported by the CDC over time. We downloaded versions of the CDC's cumulative vaccine uptake estimates that are updated retroactively as new reports of vaccinations are received on April 12, April 21, May 5, and May 26. This allowed us to examine how much the CDC's estimates of vaccine uptake for a particular day change as new reports are received. Extended

Data Fig. 3 compares the estimates of cumulative vaccine uptake for April 3-12, 2021 reported on April 12, 2021 to estimates for those same dates reported on subsequent dates. The top line shows that the cumulative vaccine uptake estimate for April 12, 2021 is, over the next month and a half, adjusted upwards by approximately 6% of the original estimate reported on April 12, 2021 . The estimate of vaccine uptake for April 11, reported on April 12, is only further adjusted upward by approximately 4% over the next 45 days. There is little apparent difference in the amount by which estimates from April 3-8 are adjusted upwards after 45 days, indicating that most of the adjustment occurred in the first 4 days after the initial report, which is consistent with the CDC's findings[13]. There is still some adjustment that occurs past day 5; after 45 additional days, estimates are adjusted upwards by an additional 2%.

There are many caveats to this analysis of CDC benchmark under-reporting, including that it depends on snapshots of data collected at inconsistent intervals, and that we mainly examine a particular window of time, April 3-12, so our results may not generalize to other windows of time. This is plausible for a number of reasons including changes to CDC reporting systems and procedures after the start of the mass vaccination program, or due to the fact that true underlying vaccine uptake is monotonically increasing over time. It is also plausible, if not likely, that the reporting delays are correlated with vaccine providers which are in turn correlated with the population receiving vaccines at a given time. As the underlying population receiving vaccines changes, so would the severity of reporting delays.

We use these results to inform our choice of benchmark uncertainty scenarios: 5% and 10%. The benchmark error is incorporated into our analysis by adjusting the benchmark estimates each day up or down by 5% or 10% (i.e. multiplying the CDC's reported estimate by 0.9, 0.95, 1.05, and 1.1). We then calculate *ddc* on each day for each error scenario, as well as for the CDC reported point estimate.

However, the benchmark data that we use here *has* been retroactively-adjusted as new reports

of vaccine administration are received, so that the scenarios we consider are in addition to the initial reporting lag which has already been accounted for. These scenarios are intended only to demonstrate the robustness of our findings to plausible latent error in the benchmark data rather than to suggest that those scenarios are at all likely. To fully account for errors in the CDC benchmark would require a close collaboration with the CDC, and to have access to its historical information and methodologies on addressing issues such as never-reporting, as occurred when reporting AIDS status[33,62].

A.3 Availability of Survey Microdata

Both Axios-Ipsos and Census Household Pulse release microdata publicly. Facebook also releases microdata to institutions that have signed Data Use Agreements. In view of the timeliness of our study, and to keep all three surveys on as equal a footing as possible, we used the aggregated results released by all three surveys rather than their microdata.

In all surveys, data collection happens over a multi-day period (or multi-week in the case of the Census Household Pulse). We calculate error for each survey wave with respect to the CDC-reported proportion of the population vaccinated up to and including the end date of each wave. Some respondents will have actually responded days (or weeks) before the date on which the estimate was released, when the true rate of vaccine uptake was lower. We use the end date instead of a mid-point as we do not have good data on how respondents are distributed over the response window. However, this means that the error we report may *underestimate* the true error in each survey, particularly those with longer fielding and reporting windows.

A.4 Census Household Pulse

The Census Household Pulse is administered by the the Bureau of Labor Statistics (BLS); the Bureau of Transportation Statistics (BTS); the Centers for Disease Control and Preven-

tion (CDC); Department of Defense (DOD); the Department of Housing and Urban Development (HUD); Maternal and Child Health Bureau (MCHB); the National Center for Education Statistics (NCES); the National Center for Health Statistics (NCHS); the National Institute for Occupational Safety and Health (NIOSH); the Social Security Administration (SSA); and the USDA Economic Research Service (ERS) (https://www.census.gov/programs-surveys/household-pulse-survey.html, visited June 5, 2021). Each wave since August 2020 fields over a 13-day time window. All data used in this analysis is publicly available on the US Census website.

The Census Household Pulse changed the question used to gauge vaccine willingness and hesitancy beginning with wave 27 (the most recent wave used in this analysis), to add a response option for respondents who are "unsure" if they will receive a COVID vaccine when they become eligible. Approximately 6.6% of all respondents reported being "unsure" in wave 27, and were coded as "vaccine hesitant" rather than "willing."

A.5 Delphi-Facebook

Facebook performs inverse propensity weighting on responses, but the reported standard errors do not include variance increases from weighting, and no estimates of design effects are released publicly. We are therefore grateful to the CMU team for providing us with estimated weekly design effects for all weeks through April 2021. The design effects are quite consistent across 2021 waves (Mean: 1.48, 95% CI: 1.48 – 1.49), so we mean-impute the design effects for May waves.

B Asymptotic Properties of *ddc*

Here we lay out the formal results underlying the interpretation of our empirical decomposition of total error into *ddc*. The first section explains how individual response behavior drives $\hat{\rho}_{Y,R}$

and sampling rate $f = n/N$. The second section describes why the relevant population size N differs between surveys of the same target population when the data collection process involves multiple processes. This clarifies the key distinction with the classic probabilistic sampling framework, and how our results are consistent with the *Law of Large Populations*[1].

B.1 The Role of Individual Response Behavior

In the Methods "Asymptotic behavior of *ddc*", we considered a logit model of the propensity score to assert that the *ddc* $\hat{\rho}_{Y,R}$ will not vanish with the population size N, regardless of how large N is. Here we provide the mathematical proof of this assertion. First, recall that the probability calculation involving Y is with respect to its finite population $\{Y_i, i = 1, \ldots, N\}$, we have $\Pr(Y = 1) = \bar{Y}_N$. Therefore, when the individual response model

$$\Pr(R = 1|Y) = \frac{e^{\alpha+\beta Y}}{1 + e^{\alpha+\beta Y}}$$

is applicable to the entire finite population (e.g., a social media platform is open to everyone, at least in theory), we have that, as $N \to \infty$, the fraction of observations

$$f \to (1 - \mu)\frac{e^{\alpha}}{1 + e^{\alpha}} + \mu\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} =: p, \tag{6}$$

where $\mu \in (0, 1)$ denotes the limit of $\bar{Y}_N$ as N increase to infinity. Here we assume such a limit exists, and it is not a trivial one (that is, μ stays away from 0 or 1). Consequently, $p \in (0, 1)$, i.e. it also stays away from 0 or 1, since it is a convex combination of $\frac{e^{\alpha}}{1+e^{\alpha}}$ and $\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$, both of which lie in $(0, 1)$. This means that we cannot make the sample n arbitrarily large (or small), such as approaching N, or even at a particular level, because it is controlled by the value of $\{\alpha, \beta\}$, which is determined by the individual response behavior (towards the specific question underlying Y).

Second, because $\mathrm{Cov}(Y, R) = \mathrm{E}(YR) - \mathrm{E}(Y)\mathrm{E}(R) = \Pr(R = 1|Y = 1)\Pr(Y = 1) - \bar{Y}_N f$, we

have

$$\hat{\rho}_{Y,R} \to \left(\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} - \frac{e^{\alpha}}{1+e^{\alpha}} \right) \frac{\sqrt{\mu(1-\mu)}}{\sqrt{p(1-p)}}. \tag{7}$$

This implies that for any given value of $\{\alpha, \beta\}$, $\hat{\rho}_{Y,R}$ will converge to a non-zero value ρ as long as $\beta \neq 0$, that is, as long as the propensity for response depends on Y itself. Consequently, the total error, relative to the standard error from simple random sampling (as a benchmark), denoted by Z,

$$Z =: \frac{\bar{Y}_N - \bar{Y}_N}{\sqrt{(1-f)\sigma^2/n}} = \hat{\rho}_{Y,R}\sqrt{N} \tag{8}$$

goes to infinity with N at the rate of $\rho\sqrt{N}$, a phenomenon that does not happen when $\beta = 0$.

B.2 Connection with the Heckman selection model

The goal of the Heckman selection model[59] is to perform estimation in the case of non-response induced by censoring a latent variable. Specifically, let each member of the population be identified via a tuple of characteristics (Y_{1i}, Y_{2i}) which satisfy:

$$\begin{aligned} Y_{1i} &= X_{1i}\beta_1 + U_{1i} \\ Y_{2i} &= X_{2i}\beta_2 + U_{2i}, \end{aligned} \tag{9}$$

where the tuples of U_i are identically and independently distributed multivariate Normal noise:

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right) \tag{10}$$

and the β_j's are regression coefficients. We seek to estimate β_1, but observe data Y_{1i} if and only if $Y_{2i} \geq 0$ (the predictors X_{ji} are observed for all members of the population, however). In our framework, the response indicator is $R_i = I(Y_{2i} \geq 0)$. The *ddc* ρ under the Heckman model (which is a theoretical model and hence this is a theoretical calculation) then is given by, using properties of the multivariate Normal,

$$\begin{aligned} \rho &= \mathrm{Corr}(Y_{1i}, I(Y_{2i} \geq 0)) \\ &= r \cdot \frac{\phi(Z_i)}{\sqrt{\Phi(-Z_i)[1-\Phi(-Z_i)]}}, \end{aligned} \tag{11}$$

where $Z_i = -X_{2i}\beta_2/\sigma_2$. Hence in this case the *ddc* is a multiplier of the correlation r in (10), where the multiplier factor λ_i' resembles the inverse Mills ratio $\phi(Z_i)/\Phi(Z_i)$, where ϕ and Φ are respectively the PDF and CDF of the standard Normal $N(0, 1)$.

Intuitively, it makes sense for ρ to be closely tied with r, since r drives the selection bias. For example, if $r = 0$, then Y_2 is independent from Y_1, and hence the sign of Y_2 will carry no information about Y_1. Therefore, for the purpose of estimating β_1, the data information is not distorted by having the sample inclusion determined by the sign of Y_2, when $r = 0$. Hence $r = 0$ must imply $\rho = 0$, and vice versa. However, r alone is insufficient to capture the impact of the biased selection mechanism, since minimally the mean of Y_2, which impacts the Z term, would influence which portion of the data is more likely to be observed. The *ddc* ρ provides a metric to capture the overall effect.

In conclusion, the *ddc* framework is closely related to the framework for inferring the population mean under the Heckman selection model (corresponding to set $X_1 = 1$). The benefit of the Heckman selection model is that we can also estimate the selection mechanism itself from the observed data thanks to the distributional assumptions about the data generating mechanism. The downside of course is that the validity of our results will depend on the reliability of the assumptions. In contrast, *ddc* makes no distributional assumptions about the data generating process, and hence it is broadly applicable. However, there is no free lunch – we cannot estimate *ddc* without external information. Nonetheless, it is a useful metric in the presence of a ground truth or plausible set of scenarios for the outcome of interest, such as in our paper.

C Additional Data Analyses

C.1 Estimates of hesitancy by demographic groups

We show estimates of our main outcomes by Education, and then by Race, in Table S1. The estimates vary by mode, but the rank ordering of a particular outcome within a single survey is roughly similar across surveys. The same estimates from Household Pulse were already presented in Table 2.

Table S1 | Levels of Vaccination, Willingness, and Hesitancy, estimated by demographic group. For each outcome, we estimate the same quantity from the three surveys. Axios-Ipsos (denoted AP). Census Household Pulse (denoted HP): wave ending March 29, 2021, n = 76,068. Delphi-Facebook (denoted FB): wave ending March 27, 2021, n = 181,949. These are the same waves as those in Table 2. Axios does not record a separate category for Asian Americans (they are lumped into "Other", so the values are left blank.

	% Vaccinated			% Willing			% Hesitant		
Education	AX	HP	FB	AX	HP	FB	AX	HP	FB
High School	28%	39%	40%	32%	40%	35%	40%	21%	25%
Some College	36	44	52	30	38	27	34	18	21
4-Year College	36	54	62	45	36	26	19	10	12
Post-Graduate	56	67	73	33	26	19	10	7	9

	% Vaccinated			% Willing			% Hesitant		
Race	AX	HP	FB	AX	HP	FB	AX	HP	FB
White	40%	50%	59%	29%	33%	24%	30%	17%	17%
Black	27	42	55	44	39	28	29	19	17
Hispanic	26	38	45	39	48	39	36	14	16
Asian		51	58		43	37		5	5

C.2 *ddc* by age / eligibility status across time

The CDC also releases vaccination rates by age groups, albeit not always in bins that overlap with the survey. For overlapping bins (seniors and non-seniors) we can calculate *ddc* specific to

each group (Extended Data Fig. 6).

The CDC only receives vaccination data for age groups from certain jurisdictions, so is likely unrepresentative of the entire US adult population. Therefore, we calculate wide bounds for what the true proportion of each age group could be based on allocating the administered doses for which we do not have age information either entirely to seniors or entirely to non-seniors. When this allocation implies a vaccination rate of more than 100% for that group, the remaining doses are allocated to the other age group. For example, if we know that on a particular day, X doses were administered to non-seniors, Y doses were administered to seniors, and Z doses were administered for which we have no age information, then the bounds for non-seniors are calculated as (X, X+Z) divided by the size of the non-senior US population. Similarly, the bounds for seniors were calculated as (Y, Y+Z) divided by the size of the senior US population.

These bounds do not incorporate any additional benchmark error, so may suffer from reporting delays or other systematic biases, and should be interpreted with caution. We do not show *ddc* for the 65+ age group due to the large width of the conservative bounds which led to unreliable estimates.

C.3 Other online polls

Clearly surveys can and do go wrong regardless of their sizes. Therefore, they key message of our analysis is *not* that "the smaller the better", but rather that (1) quality matters far more than quantity, and (2) large surveys fail more drastically than small surveys when there is non-negligible *ddc*. To highlight these points, we considered three more major online polls that ask vaccination status.

Figure 5 shows how the estimated vaccination rate of Axios-Ipsos, Data for Progress, Morning Consult, and Harris Poll tracks the CDC benchmark. The poll that is perhaps most similar to Axios-Ipsos and provides enough documentation of their methods and data, Data for Progress,

generated similar patterns as Axios-Ipsos. Their estimates tended to underestimate the vaccination rate by May, but did not suffer from overconfidence in its incorrect estimate. Data for Progress is an online-only panel run in the online vendor Lucid.

Data for Progress collects samples by the online vendor Lucid. Each wave can last up to a week and has a sample size of about $n = 1,000$. They ask:

> "As you may know, vaccines for Covid-19 have now been approved by the Food and Drug Administration and are being offered to some individuals based on specific criteria. As of today, have you been vaccinated for Covid-19?" (1) "Yes, I have received at least one Covid-19 vaccination shot," (2) "No, I have not received a Covid-19 vaccination shot."

Data for Progress' poststratification weighting weights to national numbers of "gender, age, region, education, race, the interaction of education and race, and presidential vote ([2020 presidential vote])."

Harris Poll employs an online panel with an unspecified vendor. Their weekly COVID polls are about $n = 2,000$ per wave, covering three days. They ask:

> "Which of the following best describes your mindset when it comes to getting the COVID-19 vaccine when it becomes available to you?" (1) "I plan to go the first day I am able to", (2) "Whenever I get around to it", (3) "I will wait awhile and see", (4) "I will not get a COVID-19 vaccine", and (5) "I have already received a COVID-19 vaccine."

and the analysis here only takes the last option as an indicator for vaccine uptake.

The Harris Poll weights by a propensity score by their "propensity to be online," and additionally poststratify for "age, sex, race/ethnicity, education, region, household size, employment, and household income" to population benchmarks.

Morning Consult employs their own online panel. They report a margin of 1 percentage point and a rough sample size of $n = 30,000$ per week (which corresponds to a wave). They ask:

> "Have you gotten the vaccine, or not?" (1) "Yes", (2) "No, but I will get it in the future," (3) "No, and I am not sure if I will get it in the future," and (4) "No, and I do not plan to get it."

Morning Consult weights their survey data to "a range of demographic factors, including age, race/ethnicity, gender, educational attainment, and region. State-level results were weighted separately to be representative of age, gender, race/ethnicity, education, home ownership and population density."

YouGov is also a prominent online poll. However, YouGov, unlike the other polls discussed here, investigated how their estimates track the CDC vaccination rate[63]. Therefore, we do not compare it with the other polls here. They found that the "have you been vaccinated" wording was more accurate than starting the question with "will you be vaccinated?" and including an "already" option, which tended to underestimate the vaccination rate. Their A/B test confirmed the change in question wording caused a discrepancy of about 14 percentage points even in the same poll.

YouGov's A/B test provides some indication why Harris underestimates the vaccination rate. Note that Harris, unlike Data for Progress and our three surveys in the main text, uses the wording, "when [the vaccine] becomes available to you." This is precisely the type of question wording that would underestimate vaccination rates, per YouGov. The underestimation of Morning Consult may be separately due to its questions not specifying "at least one dose," thereby

inducing a fraction of one-dose only respondents to not select "Yes." We therefore suspect the underestimation of the Harris Poll is due to the question wording rather than something systematic about online polls.

D *ddc*-based Scenario Analysis for Willingness and Hesitancy

The main quantity of interest in the surveys examined here is not uptake, but rather willingness and hesitancy to accept a vaccine when it becomes available. Our analysis of *ddc* of vaccine uptake cannot offer conclusive corrected estimates of willingness and hesitancy; however we propose *ddc*-based scenarios that suggest plausible values of willingness and hesitancy given specific hypotheses about the mechanisms driving selection bias.

D.1 Setting up scenarios

We adopt the following notation for the key random variables we wish to measure:

- V - did you receive a vaccine ("vaccination")?
- W - if no, will you receive a vaccine when available ("willingness")?
- $H = 1 - V - W$ - vaccine "hesitancy"

Just as we have studied the data quality issue for estimating the vaccine uptake, we can apply the same framework to both W and H. Unlike uptake, however, we do not have CDC benchmarks for willingness or hesitancy. We only know that $V + H + W = 1$, and therefore that

$$\mathrm{Cov}(R, V) + \mathrm{Cov}(R, H) + \mathrm{Cov}(R, W) = 0$$

Re-expressing the covariances as correlation, and recognizing that $\mathrm{Corr}(R, \cdot) = \rho_{R,\cdot}$, we obtain

$$\rho_{R,V} \cdot \sigma_V + \rho_{R,H} \cdot \sigma_H + \rho_{R,W} \cdot \sigma_W = 0$$

It is well-known that for a Bernoulli random variable, its variance is rather stable around 0.25 unless its mean is close to 0 or 1. For simplicity, we then adopt the approximation that $\sigma_V^2 \approx \sigma_H^2 \approx \sigma_W^2$. Consequently, we have

$$\rho_{R,V} + \rho_{R,H} + \rho_{R,W} \approx 0$$

As we have estimated *ddc* of vaccine uptake for each survey wave, we can further say that $\rho_{R,H}+\rho_{R,W} \approx -\hat{\rho}_{R,V}$. However, we have no information to suggest how $\rho_{R,V}$ is decomposed into *ddc* of hesitancy and willingness. Therefore, we introduce a tuning parameter, λ, that allows us to control the relative weight given to each $\rho_{R,H}$ and $\rho_{R,W}$, such that

$$-\rho_{R,H} = (1-\lambda)\hat{\rho}_{R,V}, \quad -\rho_{R,W} = \lambda\hat{\rho}_{R,V}$$

The tuning parameter λ may take on values greater than 1 and less than -1, which would indicate that the *ddc* of either willingness or hesitancy is *greater* in magnitude than that of uptake, or that selection bias is more extreme than that or vaccine uptake.

D.2 Obtaining scenario estimates

Once we postulate a particular value of *ddc*, we can use Equation 3 to solve for the population quantity of interest, say $\bar{H}_N$. Specifically, given a postulated value of $\rho_{H,R_\mathrm{w}} = r$, we can calculate $\bar{H}_N$ as follows:

$$\overline{H}_\mathrm{w} - \overline{H}_N = \underbrace{r \cdot \sqrt{\frac{N-n_\mathrm{w}}{n_\mathrm{w}}}}_{c} \cdot \sqrt{\overline{H}_N(1-\overline{H}_N)}. \tag{12}$$

Squaring both sides and rearranging, we obtain:

$$(c^2+1)\bar{H}_N^2 - (2\bar{H}_\mathrm{w} + c^2)\bar{H}_N + \bar{H}_\mathrm{w}^2 = 0, \tag{13}$$

which can be solved for $\bar{H}_N$. The two roots of the quadratic equation, which we will denote by $\{h_1, h_2\}$ with $h_1 < h_2$, corresponding $\rho_{H,R_\mathrm{w}} = r$ and $\rho_{H,R_\mathrm{w}} = -r$. Since we know the sign of r, there will be no ambiguity on which root to take.

We note that, by setting $z = r\sqrt{N}$ and rearranging (Equation 12), we have

$$\frac{\overline{H}_\mathrm{w} - \overline{H}_N}{\sqrt{\frac{1-f}{n} \cdot \overline{H}_N(1-\overline{H}_N)}} = z, \tag{14}$$

where $f = n_{\mathrm{w}}/N$. One may recognize that is the quantity for constructing the classical Wilson score confidence interval for a binomial proportion[64], but with the finite-population correction factor $(1 - f)$. This connection illuminates the meaning of the particular value of *ddc* ($\rho_{H,R_{\mathrm{w}}}$) in this context: the quantity z, which directly depends on *ddc*, is the corresponding *quantile* used in the Wilson interval. In other words, z is the multiplier or yardstick of the benchmark error (provided by simple random sampling) to measure the error in the estimator $\overline{H}_{\mathrm{w}}$. The fact that it grows with $\sqrt{N}$, when $\rho_{H,R_{\mathrm{w}}}$ does not vanish with $1/\sqrt{N}$, is precisely the explanation from the *ddc* framework.

D.3 Scenario estimates

We focus on three scenarios defined by ranges of λ that correspond to three mechanisms:

This allocation scheme allows us to pose scenarios implied by values of λ that capture three plausible mechanisms driving bias. First, if hesitant (H) and willing (W) individuals are equally under-represented ($\lambda \approx 0.5$), leading to over-representation of uptake, correcting for data quality implies that both Willingness and Hesitancy are higher than what surveys report (Extended Data Fig. 4, yellow bands). We label this the *uptake* scenario because, among the three components, uptake has the largest absolute *ddc*. Alternatively, the under-representation of the *hesitant* population could be the largest source of bias, possibly due to under-representation of people with low institutional trust who may be less likely to respond to surveys and more likely to be hesitant. This implies $\lambda \approx 0$ and is shown in the red bands. The last scenario addresses issues of *access*, where under-representation of people who are willing but not yet vaccinated is the largest source of bias, perhaps due to correlation between barriers to accessing both vaccines and online surveys (e.g., lack of internet access). This implies $\lambda \approx 1$ and upwardly corrects willingness, but does not change hesitancy.

In particular, the values used to generate the bands shown in Extended Data Fig 4 use the

following values of lambda:

- *Access* (blue bands): $\lambda \in [1, 1.2]$, and thus $\rho_W \in [-1.2\rho_V, -\rho_V]$ and $\rho_H \in [0, 0.2\rho_V]$.
- *Hesitancy* (red bands): $\lambda \in [-1.2, -1]$, and thus $\rho_H \in [-1.2\rho_V, -\rho_V]$ and $\rho_W \in [0, 0.2\rho_V]$.
- *Uptake* (yellow bands): $\lambda \in [0.4, 0.6]$ $\rho_H \in [-0.6\rho_V, -0.4\rho_V]$ and $\rho_W \in [-0.6\rho_V, -0.4\rho_V]$.

For each of the scenarios we estimates, adjustments with $\rho_{R,V}$ (*ddc* of vaccination) by each survey puts the three survey's estimates of Hesitancy and Willingness in agreement. Because the width of each band is proportional to each survey's estimated $\rho_{R,V}$ by a constant λ, it makes sense that Delphi-Facebook has the widest band and Axios-Ipsos has the narrowest band.

The *hesitancy* scenario suggests that the actual rate of hesitancy is about 31-33% in the most recent waves of Delphi-Facebook and Census Household Pulse, almost double that of original estimates. In the *uptake* scenario, both hesitancy and willingness are about 5 percentage points higher than each survey's original estimates. The *access* scenario suggests that willingness is as high as 21%, i.e. that a fifth of the US population still faced significant barriers to accessing vaccines as of late May.

Axios-Ipsos scenarios differ from those of the other two surveys due to its small *ddc*, and different question wording. The question that Axios-Ipsos uses to gauge vaccine hesitancy is worded differently from the questions used in Census Household Pulse and Delphi-Facebook. The question asks about likelihood of receiving a "first generation" COVID-19 vaccine, which may increase levels of hesitancy among respondents if they believe the survey is asking about an experimental, rather than a thoroughly tested, vaccine. We do see that Axios-Ipsos has markedly higher baseline levels of hesitancy than either Census Household Pulse or Delphi-Facebook. While this is likely driven in part by the lower estimated rates of vaccine uptake, it is also

likely due in part to question wording. Therefore, we exclude Axios-Ipsos from our scenarios of vaccine hesitancy and willingness.

The *ddc* of Axios-Ipsos is small, its estimates of hesitancy are affected less by these scenarios. Furthermore, the implied level of Hesitancy estimates for Axios-Ipsos is higher than that of the other two polls by 5-10 percentage points in the Access scenario. In fact Axios-Ipsos' *original* estimates of Hesitancy are higher than the other polls, above and beyond demographic composition differences (Table S1). This is likely to the wording of the inclusion of "first generation vaccine" in Axios-Ipsos' vaccine hesitancy question (Methods section Additional survey methodology). Because such wording differences may confound the interpretation of the scenarios (*ex ante*), we do not present Axios-Ipsos' results in the same figure as the other two surveys in the main text. To be clear, the vaccination is measured in a different question than Hesitancy (Table 1) and does not affect our presentation of vaccination-related outcomes in earlier parts of the article.

This analysis alone cannot determine which scenario is most likely, and scenarios should be validated with other studies. However, we hope that these substantive, mechanism-driven scenarios are useful for policymakers who may need to choose whether to devote scarce resources to the Willing or Hesitant populations. Extended Data Fig. 4 also shows that when positing these scenarios through a *ddc* framework, the estimates from Delphi-Facebook and Census Household Pulse disagree to a lesser extent than in the reported estimates (Extended Data Fig. 1 and 2).