Are We There Yet? Big Surveys Significantly Overestimate COVID-19 Vaccination in the US

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
[2]Department of Political Science, Stanford University
[3]Harvard College, Harvard University
[4]Department of Statistics, Harvard University
[5]Department of Mathematics, Imperial College London
[†]These authors contributed equally to this work.
[*] s.flaxman@imperial.ac.uk.

August 2021

Abstract

Accurate surveys are the primary tool for understanding public opinion towards and barriers preventing COVID-19 vaccine uptake. We compare three prominent surveys about vaccination in the US: Delphi-Facebook ($n \approx 250{,}000$ per week), Census Household Pulse ($n \approx 75{,}000$), and Axios-Ipsos ($n \approx 1{,}000$). We find that the two larger surveys are biased compared to the benchmark from the Centers for Disease Control and Prevention (CDC), and that their sample sizes lead to devastating overconfidence in those incorrect estimates. By April 26, 2021, Delphi-Facebook and Census Household Pulse estimated that at least 73% and 69% of US adults had received a first dose of COVID-19 vaccine, which was 16 and 12 percentage points higher, respectively, than the CDC's estimate (57%). Moreover, estimates of vaccine hesitancy disagree significantly between surveys – we find that these differences cannot be explained entirely by Delphi-Facebook's under-representation of racial minorities and non-college educated adults. These are examples of the Big Data Paradox[1]: when a confidence interval based on a large but biased sample exhibits both a seriously displaced center *and* a grossly underestimated width, thus leading us (confidently) away from the truth. With sufficient attention to quality control, small surveys like Axios-Ipsos can be far more reliable than large ones. We leverage a recently established data quality identity[1] to quantify sources of the estimation errors and to conduct a scenario analysis for implications on vaccine willingness and hesitancy. Our study quantifies how bias in large samples can lead to overconfidence in incorrect inferences, which is particularly problematic in studies, like those examined here, that inform high-stakes public policy decisions.

1 Which estimates should we trust?

Throughout the COVID-19 epidemic in the United States, publicly available, regularly updated, and reliable datasets have played a crucial role in informing epidemic responses at all levels of government[2] and in civil society[3]. The roll-out of vaccines across the US in 2021 has focused attention on critically important questions surrounding vaccine uptake, access, willingness and hesitancy. Policymakers and the public urgently need fine-grained spatial, temporal, and sociodemographic information about COVID-19 vaccine related attitudes and behaviors[2].

However, substantial discrepancies exist among three of the most prominent online surveys that measure vaccine-related behavior and attitudes in the US: Delphi-Facebook's COVID-19 symptom tracker[4,5] ($n \approx 250{,}000$ per week and with over 4.1 million responses in 2021), the Census Bureau's Household Pulse survey[6] ($n \approx 75{,}000$ per wave and with over 520,000 responses in 2021), and Axios-Ipsos' Coronavirus Tracker[7] (about 1,000 responses per wave, and over 10,000 responses in 2021).

Despite the large sample sizes of the first two surveys, which under standard statistical assumptions make uncertainty negligible, we observe disturbing divergences between their estimates. For example, Delphi-Facebook state-level estimates for hesitancy (defined as participants who responded that they will "definitely not" / "probably not" receive a COVID-19 vaccine, or they are unsure) from the week ending March 27, 2021 are systematically higher (2.8 percentage points on average, with standard deviation 2.9) than those from the Census Household Pulse wave ending on March 29, 2021 (Fig. 1A). The CDC has noted the discrepancies between their own reported vaccine uptake and that of the Census Household Pulse[8,9]. We observe similar discrepancies in estimates of willingness (defined as participants who responded that they will "definitely" or "probably" accept a COVID-19 vaccine, but also have not been vaccinated, Fig. 1B) and uptake (defined as people who have received at least one dose of a COVID-19

vaccine; Fig. 1C) at the state-level.

Such discrepancies can mislead, or at least confuse, policy-making. For example, the discrepancies in the estimates are large enough to render even the relative rankings of states weakly correlated (a Kendall rank correlation of 0.49, Fig. 1D-F). For instance, Missouri is the 11th most hesitant state according to Delphi-Facebook with 24.4% (95% CI: 23.0%-25.6%) of adult residents vaccine hesitant, but the Census Household Pulse estimates that only 16.7% (13.3%-20.1%) of the population is hesitant, making it their 36th most hesitant state.

These estimates also disagree with the uptake rates from the US Centers for Disease Control and Prevention (CDC). Fig. 1G-H show that in the selected March waves, on average, Delphi-Facebook and Census Household Pulse over-estimate state-level vaccine uptake by 14.8 and 8.4 percentage points, respectively. There is also barely any agreement in a survey's estimated state-level rankings with the CDC (a Kendall rank correlation of 0.26 in Fig. 1I, 0.21 in Fig. Fig. 1J). For example, Massachusetts is ranked 48th (one of the lowest) in vaccine uptake by both Delphi-Facebook and Census Household Pulse, but 7th (in the top ten) by the CDC. For context, for a state near the herd immunity threshold (70-80% based on recent estimates[10–12]), 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.

Which of these surveys can we trust? A recently proposed statistical framework[1] permits us to interrogate and quantify key sources of error in large surveys, and hence address such questions analytically. This framework has been applied to COVID case counts[13], and in other non-COVID settings[14]. Its full application requires ground-truth benchmark data, which is available for vaccine uptake because vaccine providers in the US are required to report daily vaccine inventory and distribution to the CDC[2,15]. Administrative data like these are not perfect and may themselves suffer from bias[16,17], but we conduct sensitivity analyses to ensure our results are robust to reasonable benchmark error (See Supplementary Information C.3). We are

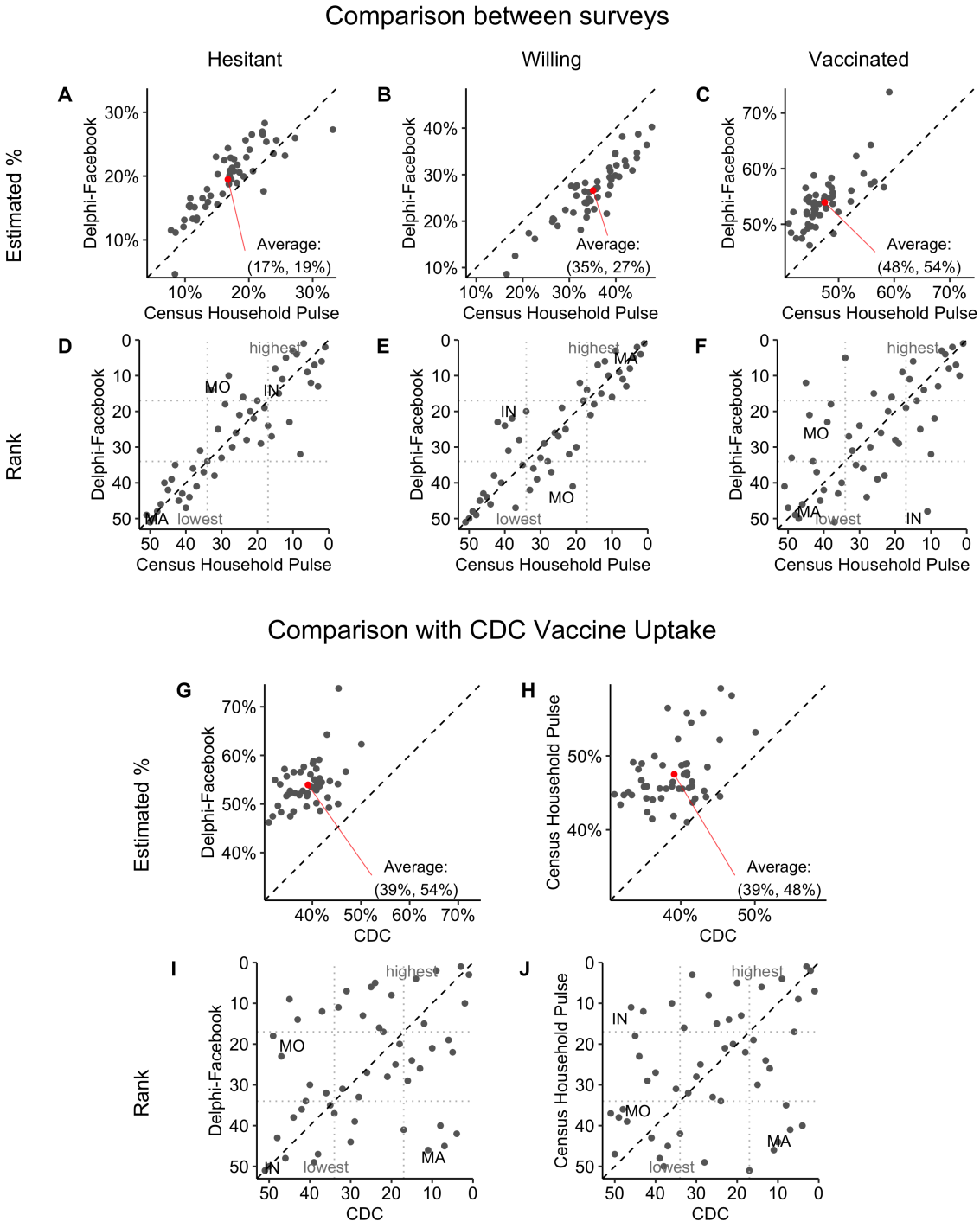

Fig. 1 | Comparisons of state-level vaccine uptake, hesitancy, and willingness across surveys and the CDC. Comparison of state-level point estimates (A-C) and rankings (D-F) for vaccine hesitancy, willingness, and uptake from Delphi-Facebook, and Census' Household Pulse. Dotted black lines show agreement 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 27, 2021. The Delphi-Facebook data is from the wave week ending March 27, 2021 and the Census Household Pulse is the wave ending March 29, 2021.

able to quantify components of estimation error driving the divergence among three surveys, apportioning it between:

- **data quality** (due to bias in coverage, sampling, measurement, response, and weighting),
- **data quantity** (driven by sample size and the weighting schemes), and
- **problem difficulty** (determined by population heterogeneity).

This assessment then allows us to borrow the magnitude of data defect observed in vaccine uptake to conduct data-driven scenario analyses for the key survey outcomes: vaccine hesitancy and willingness.

2 Conflicting estimates of vaccine uptake and the Big Data Paradox

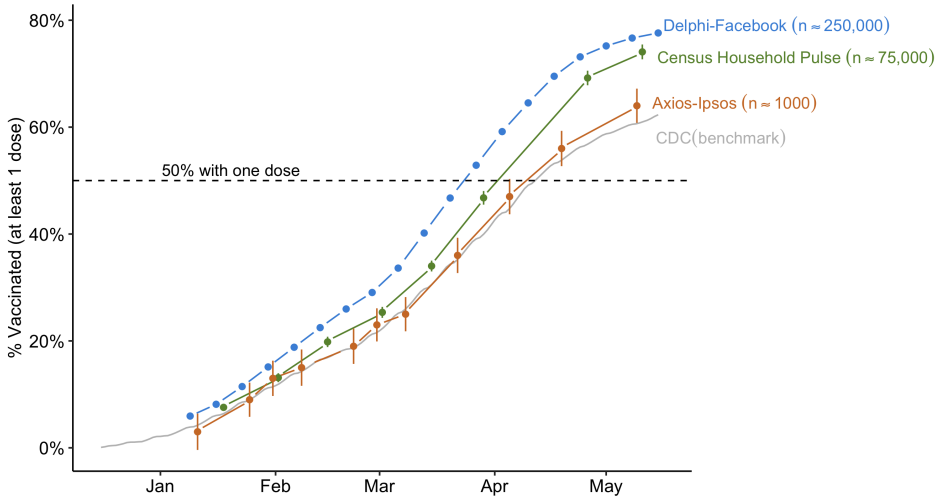

Fig. 2 | Estimates of vaccine uptake for US adults compared to CDC benchmark data, plotted by the end date (in 2021) of each survey wave. 95% confidence intervals shown are calculated based on each study's reported standard errors and design effects from weighting; the Delphi-Facebook's confidence intervals are too small to be visible.

We focus on the Delphi-Facebook, Census Household Pulse, and Axios-Ipsos surveys because they have been widely used[5,18,19], including by the US federal government[2], to understand attitudes towards COVID-19 vaccination. We focus on the national estimates from each survey because the CDC only releases vaccine uptake data that is historically-corrected for reporting delays at the national level.

Delphi-Facebook and Census Household Pulse surveys persistently overestimate vaccine uptake relative to the CDC's benchmark. For example, on April 12, the CDC reported that the vaccine uptake rate among adults reached 50% (Fig. 2). Delphi-Facebook estimates would indicate that the US passed this same milestone two weeks earlier – with a purported 52.9% (95% CI: 52.6%-53.1) rate by March 27. The Census Household Pulse wave ending on March 29 estimated the uptake rate to be 46.8% (95% CI: 45.5%-48.0%), 7 percentage points higher than the CDC's 40% rate on the same day. Despite being the smallest survey by an order of magnitude, Axios-Ipsos' estimates track well the CDC rates, and their 95% confidence intervals contain the benchmark estimate from the CDC in 10 out of 11 surveys (an empirical coverage probability of 91%).

Bias in big surveys is particularly concerning because as sample size increases, bias (rather than variance) dominates estimator error. Conventional formulas for confidence intervals mislead by conveying dire overconfidence in biased estimates. Fig. 2 shows 95% confidence intervals for vaccine uptake based on reported sampling standard errors and weighting design effects[20]. Axios-Ipsos has the widest confidence intervals, but also the smallest design effects (1.08-1.24) suggesting that its accuracy is driven more by high-quality data collection rather than post-survey adjustment. Census Household Pulse has small, but visible, 95% confidence intervals that have been greatly inflated by large design effects (4.65-4.85) indicating large weighting adjustments; however, confidence intervals still fail to include the true rate of vaccine uptake. Most concerningly, confidence intervals for Delphi-Facebook are vanishingly small – driven by

large sample size and moderate design effects (1.42-1.53) – indicating that although samples are weighted, the adjustment is not nearly enough to correct for bias in data collection.

While it is well-understood that traditional confidence intervals only capture sampling errors in surveys[21] (and not the total errors), we lack tools that would allow us to quantify nonsampling errors separately from sampling errors, and thus assess data quality. This problem is amplified in the analysis of large surveys in which the sample size pushes the sampling errors (the only part we know how to quantify in general) to a negligible level, and hence actual error is dominated by unquantified sources. These sources are responsible for the Big Data Paradox[1]: *the larger the data size, the surer we fool ourselves* when we fail to account for data quality. A large but biased sample is doubly misleading: it provides a confidence interval with an incorrect center and a grossly underestimated width. This prevents us from ever having any chance to be near the truth, and makes clear the critical importance of emphasizing data quality over data quantity.

3 Decomposing sources of error

Statistically, we can decompose the actual error into quantities capturing data quality, data quantity, and problem difficulty[1]. Given a variable of interest, Y, in a finite population of units $i = 1, \ldots, N$, of which a sample of size n is observed, where $R_i = 1$ if unit i is recorded in the sample and zero otherwise, Meng[1] shows that the error in using the sample mean $\bar{Y}_n$ to estimate the population mean $\bar{Y}_N$ can be written as

$$\underbrace{\bar{Y}_n - \bar{Y}_N}_{\text{Total Error}} = \underbrace{\hat{\rho}_{Y,R}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{Data Quantity}} \times \underbrace{\sigma_Y}_{\text{Problem Difficulty}} . \tag{1}$$

where $\hat{\rho}_{Y,R} = \text{Corr}(Y, R)$, and $\sigma_Y^2 = \text{Var}(Y)$. Note here both the correlation and variance are with respect to the finite population, $\{(Y_i, R_i), i = 1, \ldots, N\}$. Hence we use the "hat" in $\hat{\rho}_{Y,R}$ to denote that its value depends on the specific realization of $\{R_i, i = 1, \ldots, N\}$.

The "data quantity" term is intuitive: holding all else fixed, increasing the fraction of the population sampled $f = n/N$ will decrease Total Error. Similarly, the "problem difficulty" term: lower population heterogeneity (small standard deviation σ_Y of Y) results in lower estimator variance and hence lower Total Error. However, the "data quality" quantity $\hat{\rho}_{Y,R}$ is less familiar. It measures the population correlation between the outcome of interest, Y, and the indicator that a unit is observed in the sample, R. It was termed the *data defect correlation* (*ddc*) $\hat{\rho}_{Y,R} = \text{Corr}(Y, R)$ by Meng [1] because if Y is correlated with R, then some Y_i's will have a higher chance of being sampled than others, thus leading to a sample average that is a biased estimator for the population average, increasing Total Error. The *ddc* therefore is an index of data quality, and it captures both the sign and magnitude of our estimation bias.

Measuring the correlation between Y and R is not a new idea in survey statistics, nor is the observation that as sample size increases, error is dominated by bias instead of variance [22,23]. The new insight from the *ddc* framework is that $\hat{\rho}_{Y,R}$ is a general metric to index the (lack of) representativeness of the data we observe, regardless of whether the sample is obtained via a probabilistic scheme. $\hat{\rho}_{Y,R}$ is a random variable whose standard deviation is $1/\sqrt{N-1}$ *when the sampling is a simple random sample* [1]. In other words, in a simple random (probabilistic) sample, the magnitude of $\hat{\rho}_{Y,R}$ is small enough to cancel out the impact of $\sqrt{N}$ on total error, because the *ddc* goes to 0 with rate $1/\sqrt{N}$.

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 (1)). This result may be counterintuitive in the traditional survey statistics framework, which often considers how error changes as sample size n grows. The *ddc* framework considers a more general setup, which takes into account individual response behavior and its impact on sample size n.

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

$$\text{logit}[\Pr(R = 1|Y)] = \alpha + \beta Y. \tag{2}$$

This is a model for a response propensity score with a similar setup as Heckman's selection model[24]. 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.

Here, when $\beta \neq 0$, $\hat{\rho}_{Y,R}$ is determined by individual behavior, not by population size N. Hence *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. As a concrete example: a 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 be very different. 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.

This phenomenon forces us to rethink the traditional wisdom that increasing sample size necessarily improves statistical estimation, especially for large populations[1]. See Supplemental Information B.1 for a more formal exposition of this result and C.4 for a formal connection to Heckman's model.

Furthermore, identity (1) also allows us to calculate effective sample size n_{eff}, that is, the size of a simple random sample that we would expect to exhibit the same level of error as what was actually observed in a given study with a given *ddc*. Unlike the classical effective sample size[20], this quantity captures the impact of bias as well as that of variance increases from weighting and sampling. Details for this calculation are in the Supplementary Information C, where we use a weighted extension of identity (1) to incorporate weights[1].

4 Using the CDC benchmark to identify survey error

While $\hat{\rho}_{Y,R}$ 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: the sample size n of each survey wave, the estimate of vaccine uptake from each sample wave $\bar{Y}_n$, and the population size N of US adults from US Census estimates[25]. We use the CDC's report of the cumulative count of first doses administered to US adults as the benchmark $\bar{Y}_N$, and calculate $\sigma_Y = \sqrt{\bar{Y}_N\left(1-\bar{Y}_N\right)}$ because Y is binary. We apply this framework to the aggregate error observed in Fig. 2.

This method for estimating $\hat{\rho}_{Y,R}$ uses **total** error $\bar{Y}_n - \bar{Y}_N$, and thus captures not only selection bias but also any measurement bias (e.g. from question wording). However, with this estimation 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 Supplemental Information B.1)

Our analysis relies on the accuracy of the underlying CDC benchmark, which may be subject to delays and slippage in how the CDC centralizes information from states. Fortunately, the CDC updates their daily vaccination numbers retroactively as new instances of doses administered on previous days are reported to the CDC. However, 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 (Supplementary Information C.3).

The error of each survey's estimate of vaccine uptake (Fig. 3A) increases over time for all studies, most markedly for Delphi-Facebook. Problem difficulty is a population quantity that changes over time and peaks when true proportion is 50% (April 2021), then decreasing again as

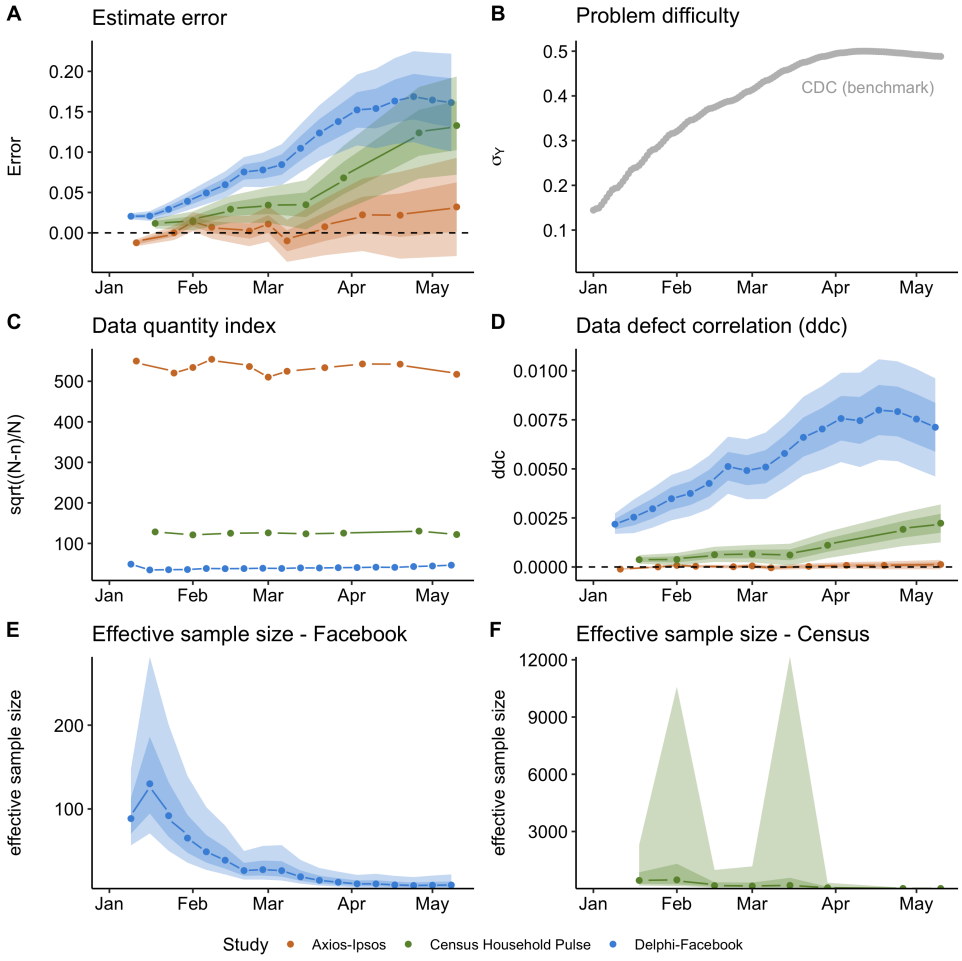

Fig. 3 | Decomposition of error in estimates of COVID-19 vaccine uptake. (A) total error $\bar{Y}_n - \bar{Y}_N$, (B) problem difficulty σ_Y, (C) an index of data quantity $\sqrt{(N-n)/n}$, and (D) the data defect correlation. Panels (E) and (F) show effective sample size accounting for bias in the Delphi-Facebook (blue) and Census Household Pulse surveys (green), respectively. Shaded bands represent actual error in scenarios of +/-5% (darker) and +/-10% (lighter) error in the CDC benchmark. (The orange ones in (A)-(D) are for the Axios-Ipsos' survey.)

the true proportion continues to rise above 50% (Fig. 3B). The data quantity index ($\sqrt{(N-n)/n}$) remains relatively constant for all studies over time, reflecting the studies' consistent raw samples sizes: about 0.1%, 0.03%, and 0.0004% of the US adult population for each wave of Delphi-Facebook, Census Household Pulse and Axios-Ipsos, respectively (Fig. 3C).

The data defect correlation, *ddc*, increases over time for Census Household Pulse and, most significantly, for Delphi-Facebook (Fig. 3D). For Axios-Ipsos, it is much smaller and steady over time, consistent with what one would expect from a representative sample. This 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. However, this does not necessarily imply a change in the response mechanism, because an identical response mechanism can result in different *ddc* as the correlation between that mechanism and the outcome changes, e.g., an individual's vaccination status Y changes over time.

A *ddc* of 0.008 (observed in Delphi-Facebook in late April) is large enough to drive effective sample size (n_{eff}) below 20, even in the scenario of 5% error in the CDC benchmark (Fig. 3E). Delphi-Facebook records about 250,000 responses per week, hence the reduction in effective sample size is over 99.9%. The maximum $\hat{\rho}_{Y,R}$ that we observe for Census Household Pulse is approximately 0.002, yet it still results in reduction in sample size of more than 99% by the same measure (Fig. 3F).

5 Comparing study designs and demographic subgroups

Sampling frames, survey modes, survey questionnaires, and weighting schemes are all instrumental to survey reliability. Table 1 compares the three surveys across these dimensions (more details available in Supplementary Information A). 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.

All three surveys are conducted online, but vary in respondent recruitment methods[26]. The Delphi-Facebook survey recruits respondents from active Facebook app users (the Facebook Active User Base, or FAUB) using daily unequal-probability stratified random sampling. 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 phone or email contact information (approximately 81% of all households on the MAF).

In comparison, Axios-Ipsos relies on inverse response propensity sampling from Ipsos' (online) KnowledgePanel, which are participants Ipsos recruits from an address-based probabilistic sample from USPS's Delivery Sequence File (DSF). 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.

All three surveys ask whether respondents have received a COVID-19 vaccine. 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[27]. However, we note that the net direction of error of this nature is unclear.

All three surveys target US adult population, but with different weighting schemes. Axios-Ipsos and Delphi-Facebook define 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.

All three surveys weight on age and gender, i.e. assign larger weights to respondents of underrepresented age-gender subgroups and smaller weights to those of overrepresented subgroups. Axios-Ipsos and Census Household Pulse also weight on education and race/ethnicity. And Axios-Ipsos additionally weights to the composition of political partisanship measured from the ABC News/Washington Post poll in 6 of the 11 waves we study. Education, a known correlate of propensity to respond to surveys[28] and social media use[29], are notably absent from Delphi-Facebook's weighting scheme, as is race/ethnicity. None of the surveys use the CDC benchmark to adjust or assess estimates of vaccine uptake. Therefore, our benchmark data provide a valid test of data quality.

Table 2 illustrates some consequences of these design choices. For education levels, Axios-Ipsos comes closest to the actual proportion of US adults even before weighting. 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 the population.

Under-representation of people without a 4-year degree likely contributed to the overestimation of overall vaccine uptake. This is because the three surveys examined here all estimate that people without a 4-year college degree are less likely to have been vaccinated compared to those with a degree (Table 2 and Supplementary Information D.1). We hasten to add that this inference from findings from these sub-populations in this survey to their population counter-

Table 1 | Comparison of designs between Axios-Ipsos, Census Household Pulse, and Delphi-Facebook studies. All surveys target the US adult population.

	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
Mode	Online, Ipsos KnowledgePanel	SMS and email to online	Facebook app to online
Length of wave	4 days, conducted weekly	2 weeks	Daily cross-section samples, reported weekly
Average size	1,000/wave	75,000/wave	250,000/week
Sampling frame	Ipsos KnowledgePanel; internet and tablets provided to about 5% of panelists who lack home internet	Census Bureau's Master Address File (individuals for whom email / phone contact information is available)	Facebook app active users
Sampling design	Inverse response propensity sampling	Systematic sample of households, adjusted for a projected response rates	Unequal-probability stratified random samples
Avg resp. rate	50%	6-8%	1%
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"
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
Weighting variables	Gender by age, race, education, Census region, metropolitan status, household income, partisanship. Partisanship weights applied 6 of 11 waves.	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 to CPS

	Composition of US Adults							Survey Estimates		
	Axios-Ipsos		Household Pulse		Delphi-Facebook		ACS	Household Pulse		
Education	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
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

	Composition of US Adults							Survey Estimates		
	Axios-Ipsos		Household Pulse		Delphi-Facebook		ACS	Household Pulse		
Race/Ethnicity	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
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.

parts requires the assumption that measured vaccination behaviors do not differ systematically between non-respondents and respondents within each education level.

The unrepresentativeness with respect to race/ethnicity and education explains part of the error of Delphi-Facebook. The racial groups that Delphi-Facebook undersamples tend to be more willing and less vaccinated. In other words, re-weighting the Delphi-Facebook survey to upweight racial minorities may bring willingness estimates closer to Household Pulse and the vaccination rate closer to CDC.

However, demographic composition alone cannot explain the discrepancies in all the outcomes. Table 2 suggests that adults without a college degree are more likely to be hesitant. Therefore reweighting Delphi-Facebook to upweight those individuals would make the overall

hesitancy estimate even higher than the original estimate, exacerbating the disagreement with Census Household Pulse. Census Household Pulse weights on both race and education and still over-estimates vaccine uptake by over ten points in late May.

Delphi-Facebook and Census Household Pulse may be unrepresentative with respect to political partisanship, which could explain some of the discrepancies. Partisanship has been found to be correlated with vaccine behavior[30,31] and with survey response[32]. However, this hypothesis is untestable because neither Delphi-Facebook nor Census Household Pulse collects partisanship of respondents; Census agencies are prohibited from asking about partisanship and political preference. Moreover, no unequivocal population benchmark for partisanship exists.

Other variables such as occupation and rurality may also contribute to the errors. However, the lack of common measurement in both the survey and population precludes full testing. We do know from CDC that there is large variation in vaccination rates by the urban-rural divide[2]. Rurality is in turn correlated with home internet access[33] which can influence the propensity to complete an online survey. 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, that is, individuals lacking access to the Internet. Through internal respondent information provided by Ipsos for their March 22 wave, we find that *dropping* these offline respondents (n = 21, or 1 percent of the sample) *increases* Axios-Ipsos' overall estimate of the vaccination rate by 0.5 percentage points, thereby increasing the total error. The 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.

Absent the full set of weighting variables and population targets, careful recruitment of

panelists is at least as important as weighting. Weighting alone cannot explain or correct the discrepancies we observe. For example, reweighting Axios-Ipsos survey data 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 with Delphi-Facebook data. The Axios-Ipsos estimate with Delphi-Facebook weighting overestimated vaccination by 2 points, whereas Delphi-Facebook overestimated by 11.

The overall takeaway is that there is no silver bullet: every small part of panel recruitment, sampling, and weighting matters for controlling the *ddc*. A *total quality control* approach — inspired by the Total Survey Error framework[34] — is a better strategy than trying to prioritize some components over others in order to improve data quality.

6 Assessing hesitancy and willingness via scenario analysis

We can leverage our knowledge of the estimation error for vaccination to provide improved estimates for hesitancy and willingness because the proportions of vaccinated (V), hesitant (H), and willing (W) individuals must sum to 1. For example, if V is an overestimate by 20 percentage points, the *under*-estimate of W and H must together sum to 20 percentage points. Without further information, we will not know how these 20 percentage points get distributed to W and H. However, we can estimate several scenarios by using *ddc* to allocate the bias in vaccine uptake to each H and W.

Specifically, the constraint $V + W + H = 1$ implies that the sum of *ddc*s of uptake (denoted ρ_V), hesitancy (ρ_H), and willingness (ρ_W), is approximately 0 (it is not exactly zero because different variables can have different variances; see Supplementary Information E). Introducing a tuning parameter λ that controls the relative weight given to bias of H and W on the *ddc* scale,

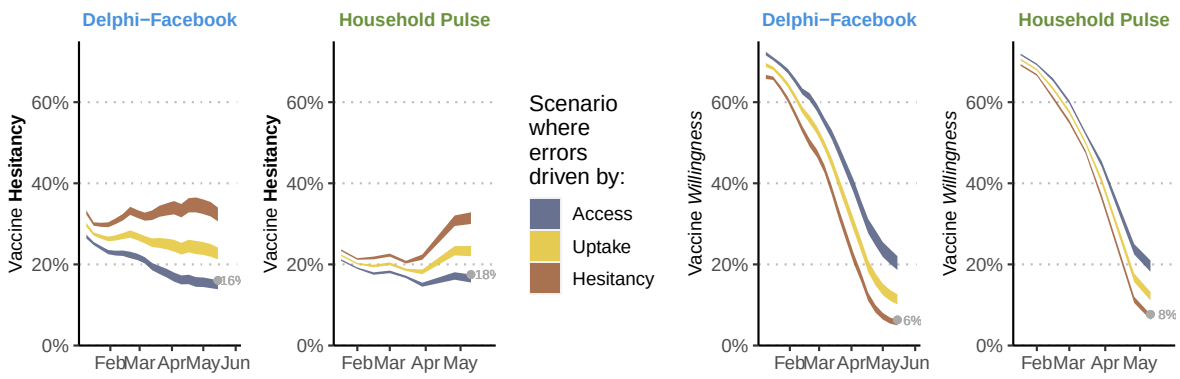

Fig. 4 | Revised estimates of hesitancy and willingness after accounting for survey errors for vaccination uptake. The gray point shows the original, reported value at the last point the time series. Each facet shows how the estimates change depending on how the error in uptake is attributed to hesitancy or willingness. The *access* scenario shows values of the outcome when $\rho_W \in [-1.2\rho_V, -\rho_V]$, and therefore $\rho_H \in [0\rho_V, 0.2\rho_V]$. The *hesitancy* scenario posits that hesitancy suffers from at least as much, if not more, bias than V, i.e. $\rho_H \in [-1.2\rho_V, -\rho_V]$ and $\rho_W \in [0\rho_V, 0.2\rho_V]$. The *uptake* scenario shows the values of the outcome when the error is split roughly equally between hesitancy and willingness, such that $\rho_H \in [-0.6\rho_V, -0.4\rho_V]$ and $\rho_W \in [-0.6\rho_V, -0.4\rho_V]$.

the zero-sum approximation implies that we can set

$$\rho_W = -\lambda\rho_V, \quad \rho_H = -(1-\lambda)\rho_V.$$

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 (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.

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. Because the *ddc* of Axios-Ipsos is small, its estimates of hesitancy are affected less by these scenarios (Supplementary Information E).

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. 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 (Fig. 1).

7 Addressing common misperceptions

The three surveys discussed in this article demonstrate a seemingly paradoxical phenomenon – the two "big surveys" are far more confident, yet also far more biased, than the smaller, more traditional Axios-Ipsos poll. As explained in Section 2 and 3, it is paradoxical only when we fall into the trap of the long-held, but incorrect, intuition that estimation errors necessarily decrease as data sizes increase[35,36].

Our proposed framework highlights the role of large population size N in determining the

devastating effects of non-negligible $\hat{\rho}_{Y,R}$ (Meng's Law of Large Populations). This may at first sound unnatural because all three surveys target the same population of US adults. However, in a multistage data collection process like those we examine here (e.g. including sampling frame definition, then sampling, followed by response), the population size that most magnifies bias in data collection is that of the earliest step at which data collection is unrepresentative; see Supplementary Information B.2 for a discussion on this somewhat subtle point. Regardless of which process (or processes) contributes most to the total bias in data collection, it is true at every stage, and overall, that data quality matters far more than data quantity. Even seemingly small defects in quality can almost completely wipe out the statistical information in our data, regardless of how large it is, as demonstrated in Section 4.

When concerns of data defect are raised, we often hear a common defense or hope that the revealed defect only affects *a particular* analysis, not necessarily other studies that use the same data. The notion of *ddc* confirms the correctness of this argument at the technical level, but also reveals its potential to mislead if it is used as the sole justification for doing business as usual. Indeed *ddc* is the correlation between a particular outcome Y and the data recording mechanism R, and hence a large *ddc* for one outcome does not imply it will be similarly large for another. However, *ddc* reveals that estimator error resulting from selection bias is merely a symptom of unrepresentativeness of the underlying sample. Selection bias tells us that respondents are not exchangeable with non-respondents, and hence it may impact *all* studies of that dataset to varying degrees. This includes study of associations[5,37] – both Delphi-Facebook and Census Household Pulse significantly overestimate the slope of vaccine uptake over time relative to that of the CDC benchmark (Fig. 2) – as well as ranking – the Census Household Pulse and Delphi-Facebook rankings are more correlated with each other (rank correlation: 0.49), than either ranking is with that of the CDC (0.21 and 0.26, respectively), as indicated in Fig. 1.

Another common response is that bias is a necessary trade-off for having data that is suffi-

ciently large for conducting high-resolution inference. Again, this is a "double-edged" argument. It is true that a key advantage of large sample sizes is that they render more data for such inference, for example about individualized treatments[38]. However, precisely because data with high-resolution is hard to come by, we often hesitate to dismiss them due to low data quality. The dramatic impact of *ddc* on the effective sample size should serve as a wake-up call to the potentially devastating overconfidence in large samples that suffer from bias, particularly in studies that can affect many people's lives and livelihoods. When a biased national sample of 250,000 contains no more usable information for estimating a national average than a simple random sample of size 20, it is a self-defeating illusion to regard the "large sample" as more useful for obtaining state-level estimates than a high-quality survey of size 1,000, for example. A highly biased estimate with a misleadingly small confidence interval can do more damage than having no estimate at all.

This is not the first time that the Big Data Paradox has reared its head, nor will it be the the last. One notable case showed how Google Trends predicted more than two times the number of doctor visits for influenza-like illnesses than did the CDC in February 2013[39]. The data collection methods of the studies we consider here have been far more carefully designed than Google Trends data, yet are still susceptible to some of the same biases. Delphi-Facebook is a widely-scrutinized survey that, to date, has been used in 10 peer-reviewed publications, including on important topics of public policy such as mitigation strategies within schools[18]. The Census Household Pulse survey 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 potential biases in data collection, but still drastically overestimate vaccine uptake. This demonstrates just how hard it is to correct for bias, even with the resources of Facebook or the US government at one's disposal, and how the impact of bias is magnified as relative sample size and sampling population size increase.

In contrast, Axios-Ipsos records only about 1,000 responses per wave, but makes a large effort to prevent selection bias (e.g., purchasing tablets for those who otherwise would be less likely to participate in an online survey). This is a telling example of why, for ensuring accuracy of inferences, data quality matters more than data quantity. Small surveys could be just as wrong as large surveys in expectation — of the three other small to medium online surveys additionally analyzed in Section D.3, two of them also miss the CDC vaccination benchmark (however, results from these other polls may be confounded by difference in question wording). 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[40] can perform better with better quality, if small, data, and it is an open question whether that approach would outperform the large, biased surveys.

There are approaches to correct for these bias in both probability and nonprobability samples alike. For COVID-19 surveys in particular, the AP-NORC multi-mode panel started to weight their COVID-19 related surveys to the CDC benchmark since June 2021, so that the weighted *ddc* for vaccine uptake is zero by design[41]. More generally, there is a growing body of literature on approaches for making inferences from data collected from nonprobability samples[40,42–44]. Other promising approaches include integrating surveys of various quality[45,46], and leveraging *ddc* from higher-quality surveys with similar outcomes to adjust for bias (as in our Scenario Analysis).

While more needs to be done to fully examine the nuances of these three surveys, we hope this first comparative study highlights the alarming implications of the *Big Data Paradox* – how large sample sizes magnify the impact of even small defects in data collection, leading to dire overconfidence in incorrect inferences.

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] 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.

[3] 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, 2021.

[4] 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.

[5] Frauke Kreuter and *et al.* Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods*, 14(2):159–163, 2020. ISSN 1864-3361. doi: 10.18148/srm/2020.v14i2.7761.

[6] JF Fields, J 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.

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

[8] 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.

[9] 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://www.cdc.gov/vaccines/imz-managers/coverage/adultvaxview/pubs-resources/sociodemographic-factors-covid19-vaccination.html.

[10] 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.

[11] Dan Lu, Alberto Aleta, Marco Ajelli, Romualdo Pastor-Satorras, Alessandro Vespignani, and Yamir Moreno. Data-driven estimate of sars-cov-2 herd immunity threshold in populations with individual contact pattern variations. https://doi.org/10.1101/2021.03.19.21253974, 2021.

[12] David Hodgson, Stefan Flasche, Mark Jit, Adam J Kucharski, CMMID COVID-19 Working

Group, et al. The potential for vaccination-induced herd immunity against the SARS-CoV-2 B.1.1.7 variant. *Eurosurveillance*, 26(20):2100428, 2021.

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

[14] 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.

[15] CDC. Trends in number of COVID-19 vaccinations, 2021. US Centers for Disease Control (CDC), `https://covid.cdc.gov/covid-data-tracker/#vaccination-trends`.

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

[17] 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.

[18] Justin Lessler, M Kate Grabowski, Kyra H Grantz, Elena Badillo-goicoechea, C Jessica E Metcalf, Andrew S Azman, and Elizabeth A Stuart. Household COVID-19 risk and in-person schooling. *Science*, 2939(April):1–11, 2021.

[19] Diane Schanzenbach and Abigail Pitts. How much has food insecurity risen? evidence from the census household pulse survey. *Institute for Policy Research Rapid Research Report*, 2020.

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

[27] Mirta Galesic, Wändi Bruine de Bruin, Jonas Dalege, Scott L Feld, Frauke Kreuter, Henrik Olsson, Drazen Prelec, Daniel L Stein, and Tamara van Der Does. Human social sensing is an untapped resource for computational social science. *Nature*, pages 1–9, 2021.

[28] An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1):1–33, 2018. doi: 10.1093/poq/nfx047.

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

[30] Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191:104254, 2020.

[31] Ariel Fridman, Rachel Gershon, and Ayelet Gneezy. COVID-19 and vaccine hesitancy: A longitudinal study. *PloS One*, 16(4):e0250123, 2021.

[32] Andrew Mercer, Arnold Lau, and Courtney Kennedy. For Weighting Online Opt-In Samples, What Matters Most? 2018.

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

[34] Paul P. Biemer and Lars E. Lyberg. *Introduction to survey quality*, volume 335. John Wiley & Sons, 2003.

[35] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[36] 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.

[37] Anna Fry, Thomas J. Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E. Allen. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):1026–1034, 2017. doi: 10.1093/aje/kwx246.

[38] Xinran Li and Xiao-Li Meng. A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance trade-off. *Journal of the American Statistical Association*, 116:353–367, 2020.

[39] 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.

[40] 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.

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

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

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

[44] Roderick JA 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.

[45] 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.

[46] 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.

[47] 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.

[48] CDC. Reporting COVID-19 vaccination demographic data, 2021. US Centers for Disease Control (CDC), https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/demographics-vaccination-data.html.

[49] 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.

[50] Emily Anthes, Madeleine Ngo, and Eileen Sullivan. Adults in all U.S. states are now eligible for vaccination, hitting Biden's target. Half have had at least one dose. *The New York Times*, 2021. https://perma.cc/7ZWP-ZBVU.

[51] Tanya Lewis. The biggest barriers to COVID vaccination for Black and Latinx people. *Scientific American*, 2021. https://perma.cc/RB5R-VGHG.

[52] Rebecca Tan and Julie Zauzmer. Vaccine sign-up in the D.C. region has been a mess. It didn't have to be this way. *The Washington Post*, 2021. https://perma.cc/D8AV-7E4R.

[53] 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.

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

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, 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 the Delphi-Facebook COVID symptom survey, and the Ipsos team for providing flags for their "offline" respondents. We thank the US Centers for Disease Control and Prevention for responding to our questions, as well as Susan Paddock and other participants at the JPSM 2021 lecture (delivered by Meng) for their comments.

Funding 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). S.F. acknowledges the support of the EPSRC (EP/V002910/1).

Author contributions V.B. conceived and formulated the research questions. All authors contributed to methodology, writing, visualization, editing, and data analysis.

Competing Interests Authors have no competing interests.

Data and materials availability All data used in this analysis is publicly available from sources listed in the references. Code and data to replicate the findings is included in our publicly available GitHub repository for this project: `https://github.com/vcbradley/ddc-vaccine-US`.

Supplementary Information

Contents

A Background of the Survey and CDC Datasets

A.1 CDC Data

The CDC benchmark data used in our analysis was downloaded from the CDC's COVID data tracker[15]. We use 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 only estimate vaccine uptake for 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. For our current purposes, 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.

The CDC does release state-level snapshots of vaccine uptake each day. These have been scraped and released publicly by Our World In Data[47]. 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.

A.2 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[7]. The PDF reports also contain Ipsos' design effects, which we have confirmed are calculated as 1 plus the variance of the (scaled) weights.

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.

A.3 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 – the Bureau of Labor Statistics (BLS); the Bureau of Transportation Statistics (BTS); the Centers for Disease Control and Prevention (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. 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 $1 + CV_{\mathrm{w}}^2$ [20], where CV_{w} is the coefficient of variation of the individual-level weights included in the microdata. 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.4 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[4]. 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.

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.

Due to data privacy constraints, the Facebook team performing the weighting does not have access to survey responses, including those for demographic questions, so is not able to use those characteristics in weighting. Similarly, the CMU team does not have access to the proprietary Facecbook data used in Facebook's weighting algorithm. Furthermore, the target population used in weighting is defined by the 2018 Current Population Survey (March Supplement), as the 2019 supplement was not yet available when the survey launched.

A.5 Availability of Microdata

Both Axios-Ipsos and Census Household Pulse release microdata publicly. Facebook also releases microdata to institutions that have signed Data Use Agreements. We are in the process of acquiring the Facebook microdata. In view of the timely nature of topics and findings, and to keep all three surveys on as equal footing as possible, in this first study 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.6 Population of Interest

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[25]. This is also what the CDC uses as the denominator in calculating rates and percentages of the US population[48].

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.

B Asymptotic Properties of *ddc* with Multi-Stage Populations

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 main text, 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

$$f \to (1 - \mu)\frac{e^{\alpha}}{1 + e^{\alpha}} + \mu\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} =: p, \tag{3}$$

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

$$\hat{\rho}_{Y,R} \to \left(\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} - \frac{e^{\alpha}}{1 + e^{\alpha}} \right) \frac{\sqrt{\mu(1-\mu)}}{\sqrt{p(1-p)}}. \tag{4}$$

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,

$$Z =: \frac{\bar{Y}_n - \bar{Y}_N}{\sqrt{(1-f)\sigma^2/n}} = \hat{\rho}_{Y,R}\sqrt{N} \tag{5}$$

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

It is important to point out that when we calculate $\hat{\rho}_{Y,R}$ via (5), we have effectively extended its meaning to capture the total survey error since $\bar{Y}_n - \bar{Y}_N$ is the actual error. Indeed, the expectation of Z^2 with respect to R (if it is random) is simply the well-known *design effect*[20]. Hence, even when $\hat{\rho}_{Y,R}$ loses its interpretation as a correlation between Y and R (such as when Y is subject to recording error), calculating it via (5) will lead to a valid measure to capture the bias from total survey error, since its square is an estimator of the design effect per population size.

B.2 Which N matters? The decomposition of population size with 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 study design. The reality is often a mix of both, with the initial outreach controlled by a sample design, but individual nonresponse

behaviors reduce the initial intended sample size. Consequently, what counts as the relevant "population size" N in the previous section depends on when and where the representativeness of our sample is destroyed, that is, 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.

Table S1 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 summarize in the main text (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), and (3) waits for individuals to respond to that survey.

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

Table S1 | Example of multi-stage population selection. 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.

As an illustration, consider the Census Pulse survey depicted in Table S1, which has three stages. This means that we can decompose the recording indicator $R = 1$ into three separate recording indicators, and *ddc* into three component *ddc*s. The first stage defines the sampling frame, which may exclude those without contact information. The second stage samples from that frame and create an *outreach list*, and the third stage is determined by individual response behaviors, that is, not everyone on the outreach list will respond. 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 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*. This is simply because the relevant population size decreases dramatically at each step.

For the Census Household Pulse in Table S1, if the 20 percent of the MAF excluded from the sampling frame (because they had no contact information) is not representative of the US Adult population, then the relevant population size is the population at Stage 1 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), then the relevant population 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 relevant population size 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[33] and our analysis of Axios-Ipsos' small "offline" population, Census Household Pulse may suffer from unrepresentativeness at Stage 1 of Table S1 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. Therefore, although asymptotics of $N \to \infty$ cannot be observed in practice and generalizing beyond the three polls is beyond the scope of this paper, the evidence presented is consistent with the Law Large Populations.

C Estimation Method for Additional Quantities

C.1 *ddc* for weighted estimators

The incorporation of survey weights into the simple *ddc* formula (Equation 1) follows standard treatment of weighted estimators and design effects in the survey literature. Given an estimator of the form $\bar{Y}_w = \sum_i w_i Y_i / \sum_i w_i$, where the subscript w on $\bar{Y}$ highlights the dependence of our estimate on weights, $\mathbf{w} = \{w_i$, for all i for which $R_i = 1\}$. (We distinguish lowercase w_i (for weights) with capital W for Willingness used in the main text.) Then it is shown in Meng[1] that

$$\bar{Y}_{\mathrm{w}} - \bar{Y}_N = \hat{\rho}_{Y,R_{\mathrm{w}}} \times \sqrt{\frac{N - n_{\mathrm{w}}}{n_{\mathrm{w}}}} \times \sigma_Y, \tag{6}$$

where $\hat{\rho}_{Y,R_{\mathrm{w}}}$ is now the population correlation between Y_i and $R_{\mathrm{w},i} = w_i R_i$ (over $i = 1, \ldots, N$). The term n_{w} is the classical "effective sample size" due to weighting[20], i.e., $n_{\mathrm{w}} = n/(1 + CV_{\mathrm{w}}^2)$, where CV_{w} is the coefficient of variation of the weights in w. The coefficient of variation is the standard deviation of weights normalized by the mean of weights. It is standard for surveys to rescale their weights to have mean 1, in which case CV_w^2 is simply $\widehat{\mathrm{Var}}(\mathbf{w})$.

C.2 Bias-adjusted effective sample size

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

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

Given a weighted estimate $\bar{Y}_{\mathrm{w}}$ with expected total mean squared error 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 mean squared error T (which is the same as variance for simple random sampling,

because its mean is an unbiased estimator for $\bar{Y}_N$). The term $E[\hat{\rho}^2_{Y,R_w}]$ 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 $\hat{\rho}^2_{Y,R_w}$ to approximate $E[\hat{\rho}^2_{Y,R_w}]$. This estimation itself is subject to error. However, it does not suffer from selection bias because our target is exactly defined by the mean of our estimator, as we aim to capture what actually has happened in this particular survey (including the impact of the weighting scheme). Hence, the only error is the sampling variability (with the caveat that the weighting scheme itself does not vary with the actual observed sample), 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.

C.3 Methods for benchmark uncertainty

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. Fig. S1 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 plot shows that the cumulative vaccine uptake for April 12, 2021 reported on that same day is adjusted upwards by approximately 6% of the original estimate over the next month and a half. 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

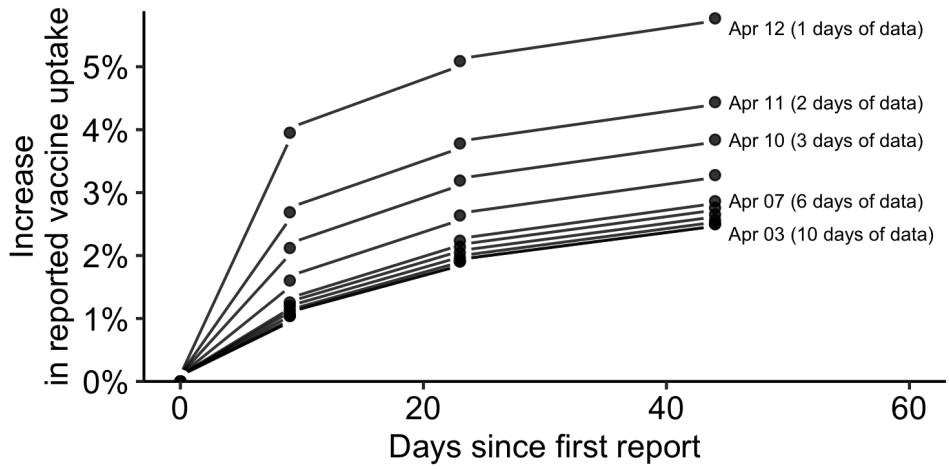

Fig. S1 | Retroactive adjustment of vaccine uptake figures for April 3-12, 2021, over the 90 days from April 12. Increase is shown as a percentage of the vaccine uptake reported on April 12.

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[15]. 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. Despite these caveats, we believe that this analysis provides reasonable guidance as to the order of magnitude that could be expected from latent systemic errors in the CDC benchmark.

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[17,49].

C.4 Connection with the Heckman selection model

The goal of the Heckman selection model[24] 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:

$$\begin{aligned} Y_{1i} &= X_{1i}\beta_1 + U_{1i} \\ Y_{2i} &= X_{2i}\beta_2 + U_{2i}, \end{aligned} \tag{7}$$

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

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right) \tag{8}$$

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,

$$\begin{aligned} \rho &= \text{Corr}(Y_{1i}, I(Y_{2i} \geq 0)) \\ &= r \cdot \frac{\phi(Z_i)}{\sqrt{\Phi(-Z_i)[1 - \Phi(-Z_i)]}}, \end{aligned} \tag{9}$$

where $Z_i = -X_{2i}\beta_2/\sigma_2$. Hence in this case the *ddc* is a multiplier of the correlation r in (8), 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.

D Additional Data Analyses

D.1 Estimates of hesitancy by demographic groups

We show estimates of our main outcomes by Education, and then by Race, in Table S2. The estimates vary by mode, but the rank ordering of a particular outcome within a single survey is roughly similar across surveys. In Table 2, we show the estimates from Household Pulse.

Table S2 | Levels of Vaccination, Willingness, and Hesitancy, estimated by demographic group. For each outcome, we estimate the same quantity from the three surveys. The Axios-Ipsos (denoted AX), Census Household Pulse (HP), and Delphi-Facebook (FB) surveys use 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

D.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 (Fig. S2). The *ddc* in the Census Household Pulse increases modestly overall over

time.

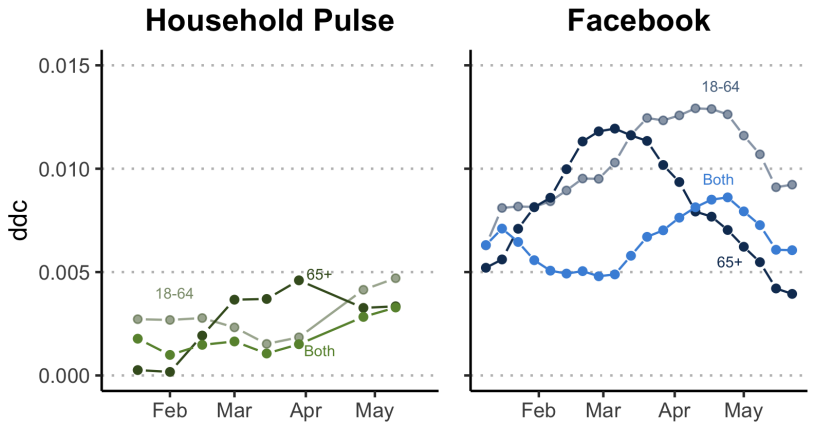

Fig. S2 | The *ddc* separated by Age Group (18-64 year-olds, and those 65 and over). Using CDC benchmark broken out by the same age bins, we recomputed separate data defect correlations (*ddc*) from weighted survey estimates (but without using a coefficient of variation adjustments). It should be noted that CDC data by demographics may not be representative of the population, due to certain jurisdictions not reporting results by demographics. The *ddc* for "Both" is computed from that same CDC data (instead of the overall benchmark shown in Figure 3).

Delphi-Facebook's *ddc* is higher overall, and shows a stark divergence between the two age groups after March 2021. The *ddc* for seniors flattens and starts to decrease after an early March peak, whereas the error rate for younger adults continues to increase through the month of March 2021, and peaks in mid-April, around the time at which all US adults became eligible[50].

This is consistent with the hypothesis that barriers to vaccine and online survey access may be driving some of the observed selection bias in Delphi-Facebook. Early in the year, vaccine demand far exceeded supply, and there were considerable barriers to access even for eligible adults, e.g., complicated online sign-up processes, ID requirements, and confusion about about cost[51,52].

A shortcoming of computing *ddc* by demographic subgroup is that the CDC benchmark data is less reliable here. They caution that

"These demographic data represent the geographic areas that contributed data and

might differ by populations prioritized within each jurisdiction's vaccination phase. Therefore, these data may not be generalizable to the entire US population."

Therefore, we do not rely on these data extensively in our main findings.

D.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 S3 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.

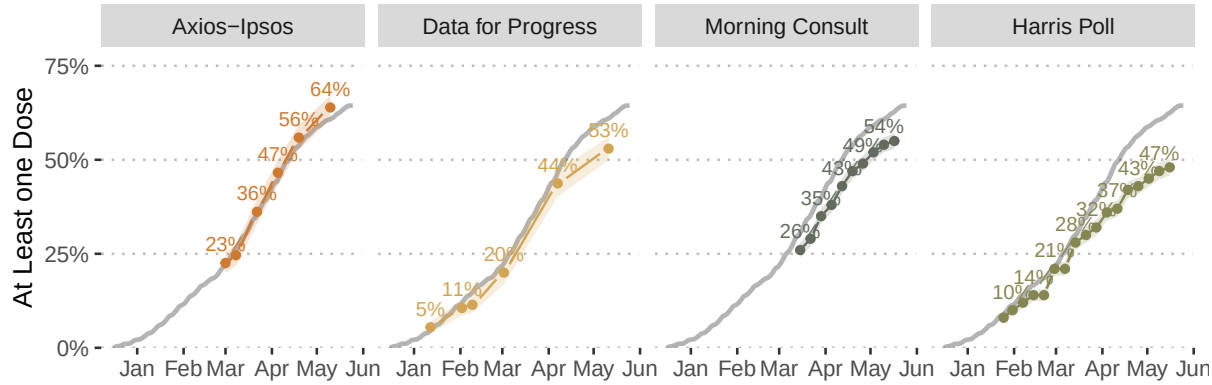

Fig. S3 | Vaccination Rates compared with CDC benchmark for four online polls. Ribbons indicate traditional 95 percent confidence intervals which are twice the nominal standard error. Gray line is the CDC benchmark.

The three surveys differed in question wording, weighting variables, and survey size. All surveys target the US adult population.

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[53]. 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.

E *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.

E.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. In particular, we focus on three scenarios defined by ranges of λ that correspond to three mechanisms as described in the main text.

E.2 Obtaining scenario estimates

Once we postulate a particular value of *ddc*, we can use identity (Equation 6) to solve for the population quantity of interest, say $\bar{H}_N$. Specifically, given a postulated value of $\rho_{H,R_\text{w}} = r$, we can calculate $\bar{H}_N$ as follows:

$$\overline{H}_\text{w} - \overline{H}_N = \underbrace{r \cdot \sqrt{\frac{N - n_\text{w}}{n_\text{w}}}}_{c} \cdot \sqrt{\overline{H}_N(1-\overline{H}_N)}. \tag{10}$$

Squaring both sides and rearranging, we obtain:

$$(c^2+1)\bar{H}_N^2 - (2\bar{H}_\text{w} + c^2)\bar{H}_N + \bar{H}_\text{w}^2 = 0, \tag{11}$$

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 $\rho_{H,R_\text{w}} = r$ and $\rho_{H,R_\text{w}} = -r$. 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 10), we have

$$\frac{\overline{H}_{\mathrm{w}} - \overline{H}_N}{\sqrt{\frac{1-f}{n} \cdot \overline{H}_N(1-\overline{H}_N)}} = z, \tag{12}$$

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[54], 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.

E.3 Scenario estimates

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 (Fig. S4). 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.

One discrepancy is that 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 S2). This is likely to the wording of the inclusion of "first generation vaccine" in Axios-Ipsos' vaccine hesitancy question (supplementary information A). 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.

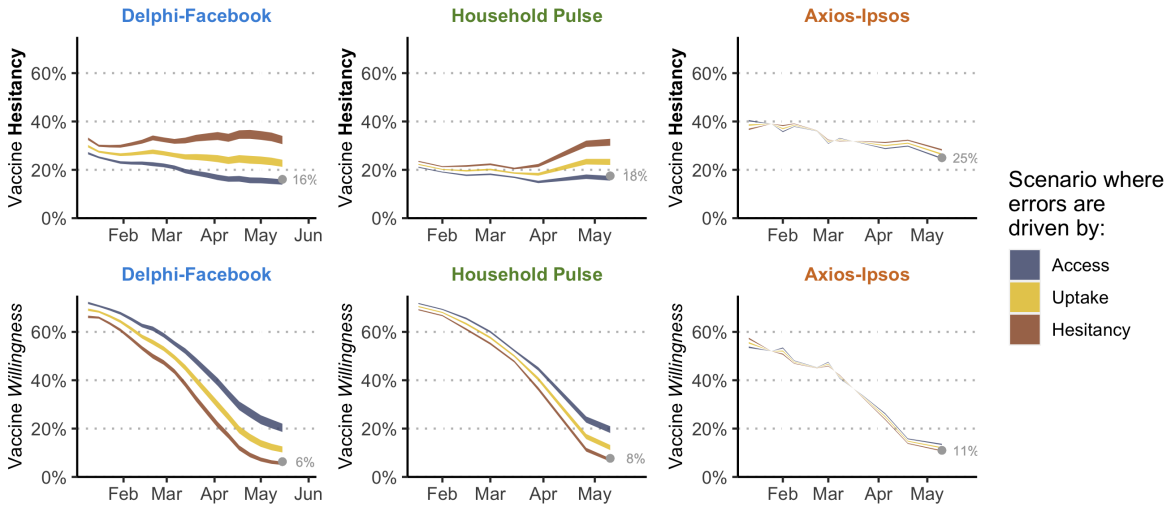

Fig. S4 | Scenario estimates for all three surveys. See Fig. 4 for full description.