Are We There Yet? Big Data Significantly Overestimates 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 Government, Harvard University
[3]Harvard College, Harvard University
[4]Department of Statistics, Harvard University
[5]Department of Mathematics, Imperial College London
[*]To whom correspondence should be addressed. s.flaxman@imperial.ac.uk.

June 2021

Abstract

Public health efforts to control the COVID-19 pandemic rely on accurate surveys. However, estimates of vaccine uptake in the US from Delphi-Facebook, Census Household Pulse, and Axios-Ipsos surveys exhibit the Big Data Paradox: the larger the survey, the further its estimate from the benchmark provided by the Centers for Disease Control and Prevention (CDC). In April 2021, Delphi-Facebook, the largest survey, overestimated vaccine uptake by 20 percentage points. Discrepancies between estimates of vaccine willingness and hesitancy, which have no benchmarks, also grow over time and cannot be explained through selection bias on traditional demographic variables alone. However, a recent framework on investigating Big Data quality (*1*) allows us to quantify contributing factors, and to provide a data quality-driven scenario analysis for vaccine willingness and hesitancy.

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 and in civil society. 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 attitudes and behaviors surrounding COVID-19 vaccines (*2*).

However, substantial discrepancies exist among three of the largest online surveys that measure vaccine-related behavior and attitudes in the US: Delphi-Facebook's COVID-19 symptom tracker (*3, 4*) ($n \approx 250,000$ per week and with over 4.1 million responses in 2021), the Census Bureau's Household Pulse survey (*5*) ($n \approx 75,000$ per wave and with over 520,000 responses in 2021), and Axios-Ipsos' Coronavirus Tracker (*6*) (about 1,000 responses per wave, and over 10,000 responses in 2021). Despite the large sample sizes of the first two surveys, rendering inconsequential error bars from traditional calculations, 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). 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 not 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

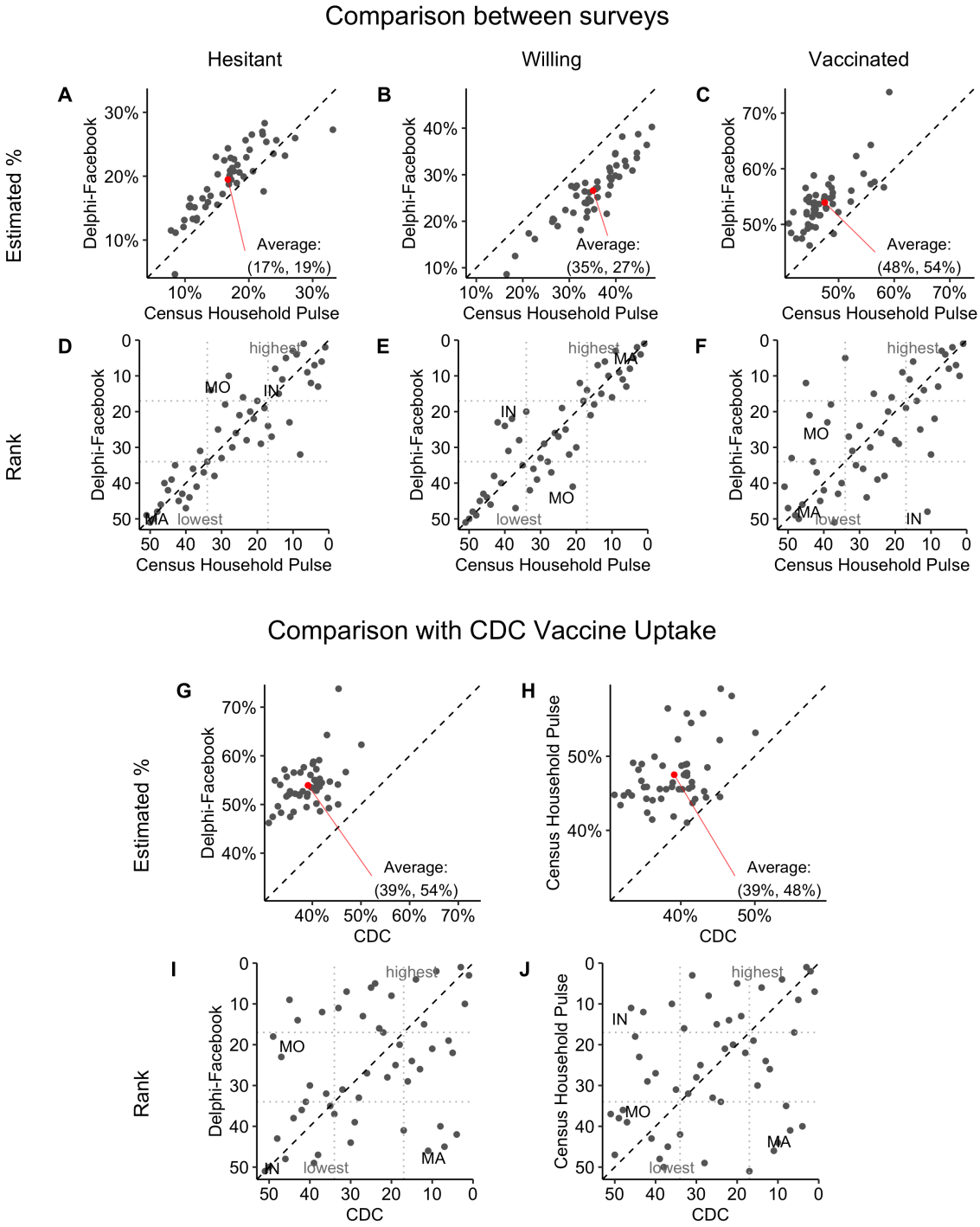

Figure 1: Comparison of state-level point estimates (A-C) and rankings (D-F) for vaccine hesitancy, willingness, and uptake from Delphi-Facebook (week ending March 27, 2021), and Census' Household Pulse (wave ending March 29, 2021). Dotted black lines show agreement and red points show average across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

discrepancies in the estimates are large enough to significantly alter the relative rankings of states by rate of vaccine hesitancy, willingness, and uptake (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 the 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, 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 little agreement in state-level rankings; for example, Massachusetts is ranked 48th in vaccine uptake by both Delphi-Facebook and Census Household Pulse, but 7th by the CDC. For context, for a state near the herd immunity threshold (70-80% based on recent estimates (*7–9*)), 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 the sources of error in big data. This framework has been applied to COVID case counts (*10*), and in other non-COVID settings (*11*). 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, 12*). We therefore are able to quantify the various components of estimation error driving the divergence among three surveys, apportioning it between data quality (due to bias in sampling, response, and weighting mechanisms), data quantity (driven by sample size and the weighting schemes), and problem difficulty (determined by population heterogeneity). This assessment then allows us to use the magnitude of data defect observed in vaccine uptake to conduct a data-driven scenario analyses for the key survey outcomes, vaccine hesitancy and willingness.

Conflicting estimates of vaccine uptake and Big Data Paradox

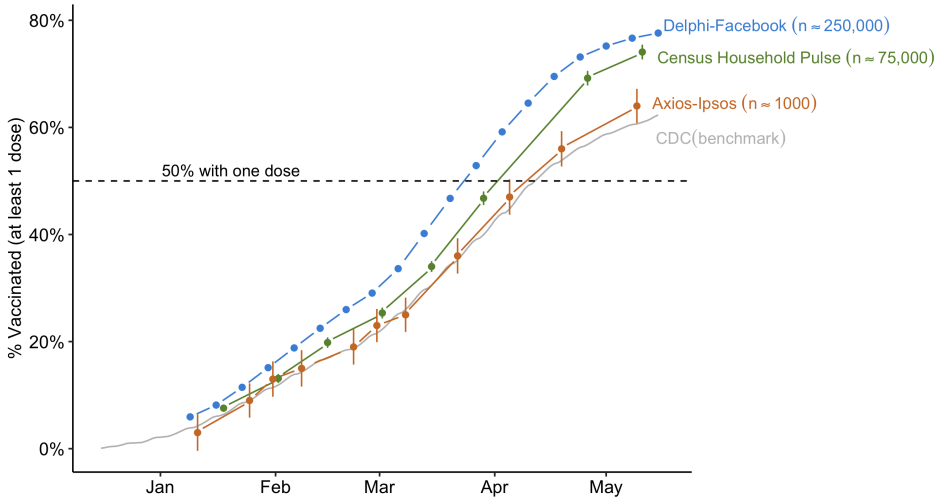

Figure 2: Estimates of vaccine uptake for US adults compared to CDC benchmark data, plotted by the end date (in 2021) of each survey wave. 95% confidence intervals shown are calculated based on each study's reported standard errors and design effects from weighting; although those for Delphi-Facebook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

We focus on the Delphi-Facebook, Census Household Pulse and Axios-Ipsos surveys because they are illustrative of surveys run by social media, governmental agencies, and survey firms, respectively. Delphi-Facebook and Census Household Pulse surveys persistently overestimate vaccine uptake relative to the CDC's benchmark. For example, on April 18, the CDC's uptake rate reached 50% (Fig. 2). Delphi-Facebook estimates would indicate that the US passed this same milestone three 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%), 8 percentage points higher than the CDC's 39% 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 over 90% of the time (10 out of 11).

The most concerning impact of biased big data is dire overconfidence. Fig. 2 shows 95% confidence intervals for vaccine uptake based on reported sampling standard errors and weighting design effects (*13*). Axios-Ipsos has the largest confidence intervals, but also the smallest design effects (1.08-1.24) suggesting that its accuracy is driven more by representativeness of the sample 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 selection bias. This is a vivid illustration of the Big Data Paradox (*1*): *the larger the data size, the surer we fool ourselves* when we fail to account for data quality. Mathematically, the probability of an incorrectly-centered confidence interval (procedure) covering the truth vanishes quickly as the sample size increases, highlighting the critical importance of emphasizing data quality over data quantity.

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, 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

$$\bar{Y}_n - \bar{Y}_N = \hat{\rho}_{Y,R} \times \sqrt{\frac{N-n}{n}} \times \sigma_Y. \tag{1}$$

It is no surprise that, holding all else fixed, increasing the fraction of the population sampled (n/N) will decrease error, or that lower population heterogeneity (small standard deviation σ_Y of Y) results in lower estimator variance and hence lower error. However, the 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. Meng (*1*) terms $\hat{\rho}_{Y,R} = \text{Cor}(Y, R)$ the *data defect correlation* (*ddc*). The *ddc* captures both the sign and magnitude of selection bias, and is therefore a measure of data quality. Studies with values of $\hat{\rho}_{Y,R}$ close to 0 indicate low (or no) selection bias for a particular outcome Y, and therefore have low estimator error.

This identity also allows us to calculate 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, n_{eff}. Unlike the classical effective sample size (*13*), this quantity captures the impact of selection bias as well as that of weighting and sampling. Details for calculating n_{eff} are in the Supplementary Materials, where we use a generalized version of Meng's identity, i.e., Equation (1), to incorporate weights.

We can apply this framework to the error in estimates of vaccine uptake from our three surveys. While $\hat{\rho}_{Y,R}$ is not directly observed, it can be easily deduced because the other terms 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 (*14*). 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.

Our analysis relies on the accuracy of the underlying CDC benchmark, which may be subject to under-reporting and other errors. 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 potential latent systemic errors, we present our results with sensitivity intervals calculated from assuming $\pm 5\%$ and $\pm 10\%$ error in the CDC's reported numbers. 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 materials B).

Fig. 3A shows that the error of each survey increases over time for all studies, most markedly for Delphi-Facebook. Problem difficulty is a population quantity that changes over time and peaks as the true proportion approaches 50% (Fig. 3B). The data quantity index ($\sqrt{(N-n)/n}$) remains relatively constant for all studies over time, reflecting the studies' consistent effective samples: 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 expects 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*, or bias in the mechanism governing which population units are observed in each sample.

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 so 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). These dramatic reductions are consequences of the *Law of Large Populations*, which we shall discuss in the concluding section.

Comparing study designs and demographic subgroups

Sampling frames, survey modes, and weighting schemes are all instrumental to survey reliability. Table 1 compares the three surveys across these dimensions (Details in supplementary materials A). All surveys are conducted online, but vary greatly in methods of respondent recruitment:

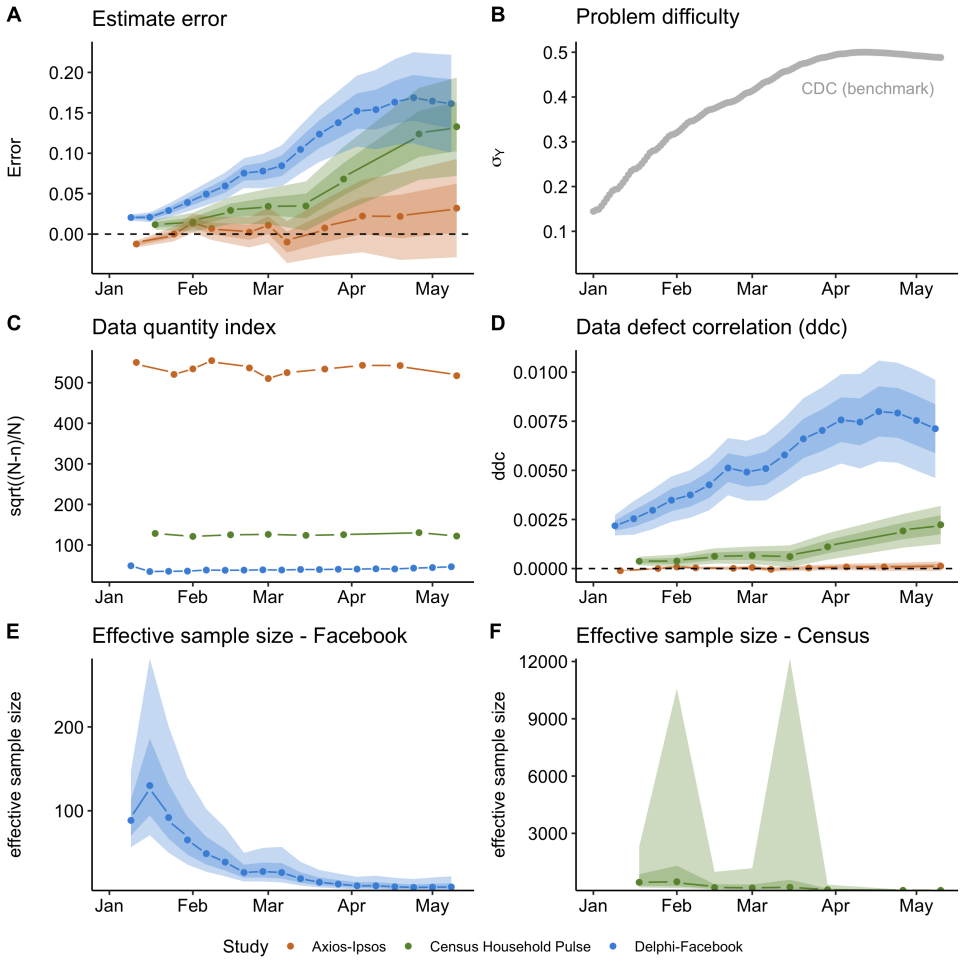

Figure 3: (A) Decomposition of error in estimates of COVID-19 vaccine uptake into actual error, (B) problem difficulty σ_Y, (C) an index of data quantity $\sqrt{(N-n)/n}$, and (D) the data defect correlation. Panels (E) and (F) show effective sample size accounting for selection bias in the Delphi-Facebook (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 red ones in (A)-(D) are for the Axios-Ipsos' survey.)

the Delphi-Facebook survey recruits respondents from active Facebook app users (the Facebook Active User Base, or FAUB), 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), and Axios-Ipsos relies on Ipsos' KnowledgePanel, an online panel recruited using an address-based probabilistic sample.

All three 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 a slightly different question, "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 studies all seek to understand the US adult population, but have 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 we show the choice between them is largely inconsequential. All studies weight on age and gender, and Axios-Ipsos and Census Household Pulse also weight on education and race/ethnicity. Education, a known correlate of propensity to respond to surveys (*15*) and social media use (*16*), as well as race/ethnicity, are notably absent from Delphi-Facebook's weighting features.

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

	Axios-Ipsos	**Census Household Pulse**	**Delphi-Facebook**
Purpose	Measure attitudes toward COVID-19	Evaluate social and economic impact of COVID-19	COVID-19 symptom surveillance
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 sample size	1,000/wave	75,000/wave	250,000/week
Sampling frame	Ipsos KnowledgePanel; panel recruited with address-based sample	Census Bureau's Master Address File	Facebook app active users
Sampling design	Inverse response propensity sampling	Systematic sample of households, adjusted for a projected response rate within each sampling area	Unequal-probability stratified random samples
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"
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 about 4 out of 10 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

Table 2 illustrates some consequences of these study designs. For education levels, Axios-Ipsos comes closest to the actual proportion of US adults even before weighting. Both Axios-Ipsos and Census Household Pulse weight on some form of education, i.e., they correct for the unrepresentativeness of the original sample with respect to education. Delphi-Facebook does not explicitly weight on education, and hence the education bias persists in their weighted estimates; those without a college degree underrepresented by nearly 20 percentage points. We observe a similar pattern with respect to 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-represent Black and Asian proportions by around 50 percent of their size in the population.

The three surveys examined here show that people without a 4-year college degree are, compared to those with a degree, both less likely to have been vaccinated and more willing to be vaccinated if a vaccine is available (Table 2). Generalizing findings from these sub-populations to the general population requires the assumption that these measured vaccination behaviors do not differ systematically between non-respondents and respondents, within each education level. If people with lower educational attainment are *under*-represented in the survey, the survey will suffer from an *over*-estimation of vaccine uptake.

The unrepresentativeness with respect to race/ethnicity and education explains part of the discrepancy in outcomes. 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 all of the discrepancies. Census Household Pulse weights on both ethnicity and education and still over-estimates vaccine uptake by a considerable margin in late May. But adults without a college degree are also more

	Composition of U.S. 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 U.S. 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.

likely to be hesitant, and Delphi-Facebook also undersamples that group. Reweighting Delphi-Facebook by education would make its Hesitancy estimate even higher than the original estimate, exacerbating the disagreement with Census Household Pulse.

Therefore, other variables, such as occupation and rurality, may contribute to the differences in estimates, but we are unable to directly examine them because they are either not reported in the surveys or no population benchmark exists. However, we do know from CDC that there is large variation in vaccination rates by rurality (*2*), which is known to be correlated with home internet access (*17*), an important factor influencing 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. Analysis of age-group-level *ddc* (see supplementary materials D.1) further suggests that selection bias in Delphi-Facebook may be correlated with the relative timing in which different age groups became eligible for the vaccine.

Delphi-Facebook and Census Household Pulse may also be non-representative with respect to political partisanship, which has been found to be correlated strongly with vaccine behavior (*18, 19*). Axios-Ipsos incorporates political partisanship in their weighting for about 40% of their waves, but neither Delphi-Facebook nor Census Household Pulse collects partisanship of respondents.

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. Naively, one might derive "corrected" estimates of W and H by increasing each raw estimate by 10 percentage points. However, we can improve upon this approach by using *ddc* to instead allocate the *selection bias* in vaccine uptake to each H and W.

As we show in supplementary materials E, the constraint $V + W + H = 1$ implies that the sum of *ddc*s of uptake, hesitancy, and willingness (denoted by ρ_V, ρ_H and ρ_W, respectively) is approximately 0 (it is not exactly zero because different variables can have different variances). Introducing a tuning parameter λ that controls the relative weight given to selection bias of H and W on the *ddc* scale, 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 selection bias. First, if hesitant (H) and willing (W) individuals are equally under-sampled ($\lambda \approx 0.5$), 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 selection bias, possibly due to under-representation of people with low institutional trust who may be less likely to respond to surveys and more likely to be hesitant. This implies $\lambda \approx 0$ and is shown in the red bands. The last scenario addresses issues of **access**, where under-representation of people who are willing but not yet vaccinated is the largest source of bias, perhaps due to correlation between barriers to accessing both vaccines and online surveys (e.g., lack of internet access). This implies $\lambda \approx 1$ and upwardly corrects willingness, but does not change hesitancy.

In the most recent waves of Delphi-Facebook and Census Household Pulse, the **hesitancy** scenario suggests that the actual rate of hesitancy is about 31-33%, almost double that of original estimates. In the **uptake** scenario, both hesitancy and willingness increase by about 5 percentage points. In the **access** scenario, the proportion of the adult population that is willing increases from 7-8% to about 21%, tripling in size, and suggesting that almost one fifth of the US population still faces significant barriers to accessing vaccines.

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).

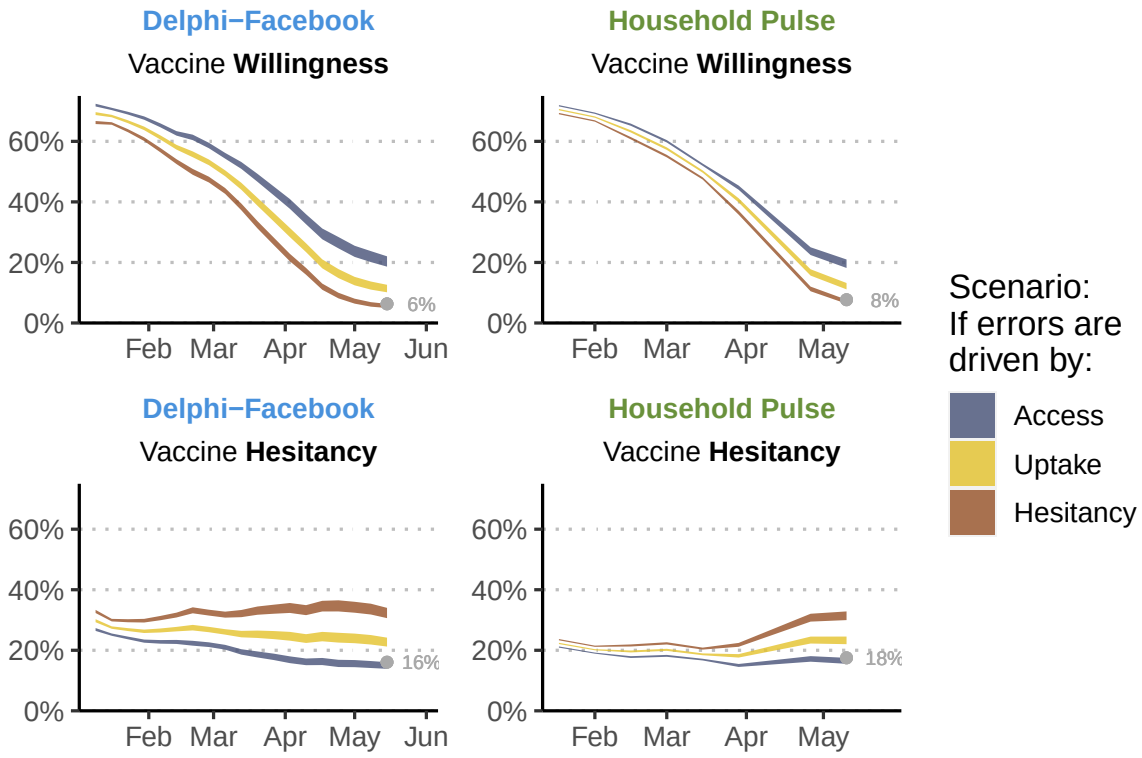

Figure 4: Revised estimates of hesitancy and willingness after accounting for survey errors for vaccination uptake. The gray point shows the original, reported value at the last point the time series. Each facet shows how the estimates change depending on how the error in uptake is attributed to hesitancy 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, selection 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 $\rho_H \in [-0.6\rho_V, -0.4\rho_V], \rho_W \in [-0.6\rho_V, -0.4\rho_V]$.

Understanding the *Law of Large Populations* and its consequences

The three surveys discussed in this article demonstrate a seemingly paradoxical phenomenon, that is, the accuracy of our results decreases with the survey size. It is paradoxical because of our long-held intuition that estimation errors decrease with data sizes. However, as proved mathematically in (*1*), this intuition only applies to probabilistic samples in which the *ddc* ($\hat{\rho}_{Y,R}$) is vanishingly small. More precisely, the value of the right-hand side of identity in Equation (1) depends on population size N (through the $\sqrt{N-n}$ term), and n is typically tiny compared to N, hence, to keep error from exploding in large populations, $\hat{\rho}_{Y,R}$ must compensate by decreasing at a rate of $1/\sqrt{N}$.

Probabilistic samples ensure this control, while non-probabalistic samples, such as those based on self-reported big data, have no such guarantee. Instead, response propensity is driven by individual behaviors, which tend to be correlated with the answers we care about, for example, because people with strong opinions may tend to use social media more to express them. Consequently, there is no reason for *ddc* to depend on population size, let alone to vanish at the rate of $1/\sqrt{N}$. The *ddc* can also be non-vanishing in probabilistic samples, such as Census Household Pulse, that suffer from nonresponse bias that is correlated with outcomes. In such cases, for a given *ddc*, the error rate goes up with $\sqrt{M}$, where M is the original intended sample size. This is because respondents can respond only if sampled, so the original sample is the population from which respondents are drawn.

This phenomenon – when *ddc* is not fully controlled by probabilistic sampling and thus the error rate increases with the square root of the (parental) population size – constitutes the *Law of Large Populations*. Meng (*1*) establishes this law and illustrates it empirically using 2016 US election polling data, and we further demonstrate this phenomenon in the context of COVID-19

vaccine uptake in the US. This law highlights the critical role of *ddc* and explains the Big Data Paradox. Resulting insights can help us to build new and better intuition for dealing with data suffering from selection bias, especially in the case of Big Data.

For example, when concerns of selection bias in data are raised, we often hear a common defense or hope that the revealed selection bias only affects *that* study, 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 potentially misleading nature 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, as captured by the R-mechanism. Selection bias tells us that respondents are not exchangeable with non-respondents, and hence it may impact *all* studies to varying degrees. This includes study of associations (*4, 20*): 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 ($\rho = 0.64$), than either ranking is with that of the CDC (0.31 and 0.33, respectively), as indicated in Fig. 1.

Another common response is that bias is a necessary trade-off for having data that is sufficiently large for conducting high-resolution inference. Again, this is a "double-edged" argument. It is very true that a key advantage of Big Data is that it renders more data for such inference, such as about individualized treatments (*21*). However, precisely because data with high-resolution is hard to come by, we tend to be very reluctant to discount them due to low data quality. The dramatic impact of *ddc* on the effective sample size should serve as a wake-up call to our potentially devastating overconfidence in biased Big Data, particularly in studies that can affect

many people's lives and livelihoods.

This is not the first time that the Big Data Paradox has reared its head, nor the last time that it will. One notable example is that of Lazer et al. (2014), which examines how Google Trends predicted more than two times the number of doctor visits for influenza-like illnesses than did the CDC in February 2013 (*22*). 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 6 peer-reviewed publications, most recently in *Science* (*23*). 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 selection bias, even with enormous sample sizes and the resources of Facebook or the US government at one's disposal.

In contrast, Axios-Ipsos records only about 1,000 responses per wave and is likely too small to make reliable inferences for sub-national geographies, but makes more of an effort to prevent selection bias for national estimates (e.g., their effort of 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 far more than data quantity, and therefore that investing in data quality (particularly in sampling, but also in weighting) is wiser than relying on data quantity. While much 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 *Law of Large Populations* – the mathematically proven fact that compensating for low data quality by increasing data quantity is a losing strategy.

References

1. X.-L. 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**, 685 (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* `https://perma.cc/FV7A-USGC`.

3. N. Barkay, *et al.*, `https://arxiv.org/abs/2009.14675` (2020).

4. F. Kreuter, *et al.*, Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods* **14**, 159 (2020).

5. J. Fields, *et al.*, Design and operation of the 2020 Household Pulse survey. (2020). U.S. Census Bureau. `https://perma.cc/JC3D-3LBY`.

6. C. Jackson, M. Newall, J. Yi, Axios Ipsos Coronavirus Index (2021). `https://www.ipsos.com/en-us/news-polls/axios-ipsos-coronavirus-index`.

7. E. J. Haas, *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).

8. D. Lu, *et al.*, `https://doi.org/10.1101/2021.03.19.21253974` (2021).

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

10. W. Dempsey, `https://arxiv.org/abs/2005.10425` (2020).

11. M. Isakov, S. Kuriwaki, Towards principled unskewing: Viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review* **2** (2020).

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

13. L. Kish, *Survey Sampling* (1965).

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

15. An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly* **82**, 1 (2018).

16. B. Auxier, M. Anderson, Social media use in 2021. *Pew Research Center* (2021).

17. C. Ryan, Computer and internet use in the United States: 2016. *American Community Survey Reports* **ACS-39** (U.S. Census Bureau, Washington, DC, 2017).

18. H. Allcott, *et al.*, Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics* **191**, 104254 (2020).

19. A. Fridman, R. Gershon, A. Gneezy, COVID-19 and vaccine hesitancy: A longitudinal study. *PloS One* **16**, e0250123 (2021).

20. A. Fry, *et al.*, Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology* **186**, 1026 (2017).

21. X. Li, X.-L. 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 (2020).

22. D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google Flu: Traps in big data analysis. *Science* **343**, 1203 (2014).

23. J. Lessler, *et al.*, Household COVID-19 risk and in-person schooling. *Science* **2939**, 1 (2021).

24. C. Appel, *et al.*, Data on COVID-19 (coronavirus) by Our World in Data (2021). `https://github.com/owid/covid-19-data`.

25. 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`.

26. X. M. Tu, X.-L. Meng, M. Pagano, The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* **88**, 26 (1993).

27. P. Bouman, V. Dukic, X.-L. Meng, A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica* pp. 325–357 (2005).

28. E. Anthes, M. Ngo, E. 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`.

29. T. Lewis, The biggest barriers to COVID vaccination for Black and Latinx people. *Scientific American* (2021). `https://perma.cc/RB5R-VGHG`.

30. R. Tan, J. 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.

31. L. D. Brown, T. T. Cai, A. DasGupta, Interval estimation for a binomial proportion. *Statistical Science* **16**, 101 (2001).

Acknowledgments

We thank 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. We thank the US Centers for Disease Control and Prevention for responding to our questions.

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.

List of Supplementary Materials

Supplementary Materials

A Background materials on the four data sets studied

A.1 CDC Data

The CDC benchmark data used in our analysis was downloaded from the CDC's COVID data tracker (*12*). 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 (*24*). 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 (*6*). 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 be confusing to respondents. We 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 $1 + CV_w^2$ (*13*), 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 (*3*). 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 Data Resolution

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 their 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 (*14*). This is also what the CDC uses as the denominator in calculating rates and percentages of the US population (*25*).

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 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 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 (*12*). 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

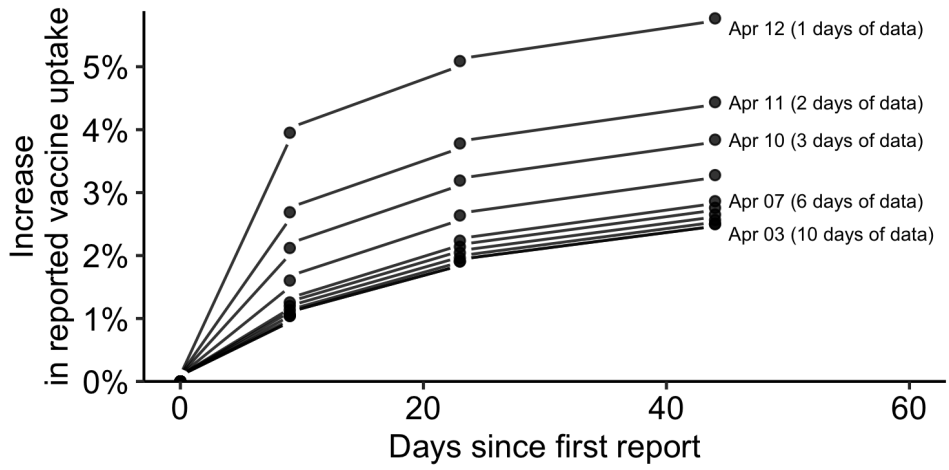

Figure S1: Retroactive adjustment of vaccine uptake figures for April 3-12, 2021, over the 90 days from April 12. Increase is shown as a percentage of the vaccine uptake reported on April 12.

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 truly 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 (*26, 27*).

C Formulas for Additional Quantities

C.1 *ddc* for weighted estimators

To calculate *ddc* in this analysis, we use an extension from Meng (*1*) that incorporates sampling weights. 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. The generalized identity is given by:

$$\bar{Y}_{\mathrm{w}} - \bar{Y}_N = \hat{\rho}_{Y,R_{\mathrm{w}}} \times \sqrt{\frac{N - n_{\mathrm{w}}}{n_{\mathrm{w}}}} \times \sigma_Y \tag{2}$$

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 (*13*), i.e., $n_{\mathrm{w}} = n/(1+CV_{\mathrm{w}}^2)$, where CV_{w} is the coefficient of variation of the weights in w (not to be confused with willingness W), as defined above.

C.2 Bias-adjusted effective sample size

Meng (2018) derives the following formula for calculating a bias-adjusted *effective sample size*, or n_{eff}:

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

Given 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 expected on average from a particular recording mechanism R and a chosen weighting scheme.

Following (*1*), 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.

D Estimates of Hesitancy and Willingness by Demographic Groups

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

Table S1: 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 (AX), Census Household Pulse (HP), and Delphi-Facebook (FB) surveys use the same waves as those in Table 2.

	% 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.1 Separate *ddc* estimates by Age Group

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.

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 (*28*).

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

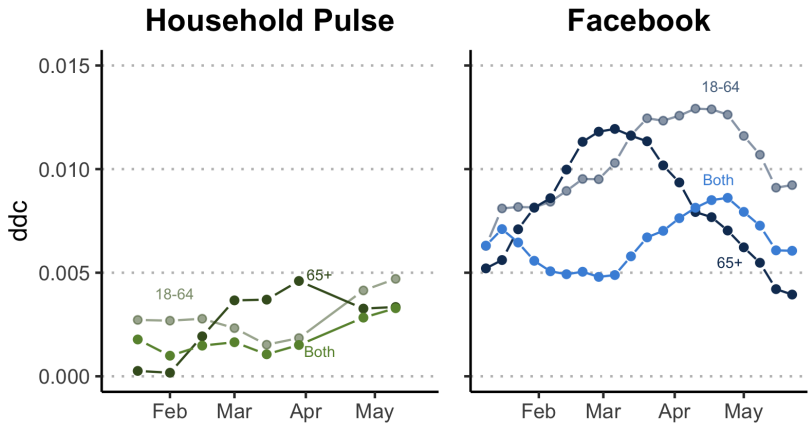

Figure S2: The *ddc* separated by Age Group (18-64 year-olds, and those 65 and over). Using CDC benchmark broken out by the same age bins, we recomputed separate data defect correlations (*ddc*) from weighted survey estimates (but without using a coefficient of variation adjustments). It should be noted 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).

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 (*29, 30*).

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.

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, w, 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 2) 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:

$$\overline{H}_\text{w} - \overline{H}_N = \underbrace{r \cdot \sqrt{\frac{N-n_\text{w}}{n_\text{w}}}}_{c} \cdot \sqrt{\overline{H}_N(1-\overline{H}_N)}. \tag{3}$$

Squaring both sides and rearranging, we obtain:

$$(c^2+1)\bar{H}_N^2 - (2\bar{H}_\text{w} + c^2)\bar{H}_N + \bar{H}_\text{w}^2 = 0, \tag{4}$$

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 3), we have

$$\frac{\overline{H}_{\mathrm{w}} - \overline{H}_N}{\sqrt{\frac{1-f}{n} \cdot \overline{H}_N(1 - \overline{H}_N)}} = z, \tag{5}$$

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 (*31*), 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 exactly the explanation underlying Law of Large Populations (*1*).