Does voting by mail increase fraud? Estimating the change in reported voter fraud when states switch to elections by mail.

Jonathan Auerbach and Steve Pierson
Technical Report*
Office of Science Policy
American Statistical Association

October 26, 2020

Abstract

We estimate the change in the reported number of voter fraud cases when states switch to conducting elections by mail. We consider two types of states: states where a large number of voters receive their ballots by mail (receive-by-mail states, RBM) and a subset of these states where all registered voters are automatically sent ballots by mail (vote-by-mail states, VBM). We then compare the number of voter fraud cases in RBM (VBM) states to the number of cases in non-RBM (non-VBM) states, using two approaches standard in the social sciences. We find no evidence that voting by mail increases the risk of voter fraud overall. Since 2016, RBM (VBM) states have reported similar fraud rates to non-RBM (non-VBM) states. Moreover, we estimate Washington would have reported eighty *more* cases of fraud had it not introduced its VBM law in 2011. While our analysis of the data considers only two of many possible approaches, we argue our findings are unlikely were fraud more common when elections are held by mail.

Introduction

U.S. voters are currently participating in the 2020 general election, which will determine the next president as well as other public officials at the federal, state, and local level. While Election Day is officially Tuesday, November 3rd, many voters are casting their ballots early—either in person or by mail. This technical report examines the claim that states can expect more cases of voter fraud when ballots are distributed by mail. It does not consider the consequences of early in-person voting or other challenges facing mail-in voting, such as the timeliness of the U.S. Postal Service or the reporting of election results. Nor does it consider whether voters are more likely to incorrectly mark their ballots or whether election workers are more likely to incorrectly reject ballots.

Absentee voting—voting outside an assigned voting station—began during the Civil War, when Union soldiers cast their vote for the 1864 general election from military camps. Over the past two decades, absentee voting by mail has increased markedly, from one-tenth of all votes in the general election to nearly one-quarter[1] (also see the Current Population Survey). All fifty states now allow voters to request a ballot by mail, although sixteen require a valid excuse. Five states automatically mail ballots to registered voters: Colorado, Hawaii, Oregon, Utah, and Washington.[2]

Many states have temporarily expanded voting by mail in light of the coronavirus pandemic, and the 2020 general election may be the first presidential election where voting by mail is the dominant form of participation. Eighty million voters are predicted to receive their ballots by mail,[3] fifty-eight percent of all

*This report was written by researchers at the American Statistical Association to evaluate policy issues of statistical importance. This report is intended for research purposes only and is not an official position, statement, or policy of the American Statistical Association on these issues. We thank Avi Feller, Karen Kafadar, and Philip Stark for helpful comments. The authors can be contacted by email at jonathan@amstat.org.

votes cast in the 2016 election. That is nearly triple the twenty-nine million voters that received their ballots by mail in the 2018 general election—itself twenty-eight percent of all 2018 ballots, up from twenty-four percent in 2016.[1]

The expansion of voting by mail has rekindled the long-standing debate between proponents of election security and accessibility. Vote-by-mail policies, and absentee voting in general, are being reviewed by legislatures and courts around the country.[4–6] A central question is whether voting by mail in the 2020 election will result in additional cases of voter fraud. That is, whether fewer cases of voter fraud would be reported if states limit voting by mail.

Studies suggest voting by mail increases political participation—to the advantage of neither Democrats nor Republicans[7,8]—although young and minority voters are more likely to have their ballots rejected.[9,10] However, despite a large literature on voter fraud,[11,12] few have investigated the effect of expanding voting by mail on voter fraud. Spakovsky (2020) argues conducting elections by mail instead of in person increases the opportunity for fraud, citing the Heritage Foundation election fraud database.[13] Others counter that even if fraud in vote-by-mail elections is more frequent than in-person voting, fraud overall amounts to a trivial percent of all ballots cast in an election,[14–17] assuming most fraud cases are reported. They also point to additional safeguards, such as risk limiting audits, which review all modes of voting to help ensure that the reported outcome of an election is correct.[18–20] Such audits have been completed successfully in vote-by-mail states, for example, the 2018 Colorado elections.

We found no study that examines the increase in voter fraud that can be expected when states switch to conducting elections by mail—even though such increases are the stated reason for many absentee-voting restrictions. We believe meaningful progress towards answering this question can be achieved by carefully comparing states with vote-by-mail elections to states without them. This technical report considers two methods, standard in the social sciences, for making these comparisons. It examines the number of criminal investigations for voter fraud in two types of mail-in states: states where a large number of voters receive their ballots by mail (receive-by-mail states, RBM) and the subset of these states where registered voters are automatically sent ballots by mail (vote-by-mail states, VBM). These numbers are then compared to the number of cases in the remaining non-RBM (or non-VBM) states using differences-in-differences and penalized regression. See 21, 22, and 23 for a discussion of these and related approaches.

After making these comparisons, we find no evidence that voting by mail increases the risk of voter fraud overall; if voting by mail creates more opportunities for fraud, those opportunities do not appear to have been realized in the data. We report the details of our analysis in four sections. Section 1 reviews how states distribute ballots by mail, motivating how fair comparisons might best be made between states that hold elections by mail and states that do not. Section 2 examines the difference in the annual number of cases per eligible voter between 2000-2004 and 2016-2020 for both RBM (VBM) states and non-RBM (non-VBM) states. Section 3 combines fraud cases from non-VBM states to best represent what select VBM states might have looked like had they not passed a VBM law, using the matrix completion approach of 24. We conclude with a discussion of our findings and important limitations.

Section 1. Vote by mail refers to how voters receive their ballots; most voters that "vote by mail" actually return their ballots in person

We begin by reviewing the different ways states distribute ballots by mail. Colorado, Oregon, Utah, and Washington automatically mailed ballots to registered voters in the 2016 and 2018 general elections. These states represent fifteen million eligible voters, 6.3 percent of eligible voters countrywide, according to the United States Election Project[25] We refer to these states as vote-by-mail states (VBM). VBM states have higher rates of voter turnout than the U.S. overall. In 2018, fifty-eight percent of eligible VBM voters actually voted, 7.2 percent of all voters.

In addition, California, New Jersey, Hawaii, Nevada, Vermont and Washington D.C. are automatically mailing ballots to registered voters in the 2020 general election (representing an additional thirty-six million eligible

voters, fifteen percent of eligible voters countrywide). This analysis does not consider these states to be VBM states since they did not mail ballots to all registered voters in past general elections.

Arizona, California, Hawaii, and Montana mailed ballots to a large number of eligible voters in the 2016 and 2018 general elections, according to the Election Administration and Voting Surveys. These states represent thirty-three million eligible voters, fourteen percent of eligible voters countrywide. We refer to these four states, plus the four VBM states, as receive-by-mail states (RBM). Note that a majority of RBM (VBM) voters do not return their ballots by mail; most return their ballots to designated locations in person.[26,27] Thus, RBM (VBM) refers to how ballots are distributed, not returned.

We further summarize RBM states in four figures, Figures 1.a through 1.d, included in the appendix. All figures in this report use state-level general election data from the U.S. Election Assistance Commission Election Administration and Voting Surveys (EAVS) and the United States Election Project (EP). The Election Project records the number of eligible voters and the number of actual voters by state since 2000. The EAVS provides additional information on the administration of general elections since 2004, although the data are not reported for all states and years.

We use fraud cases reported in the News21 Election Fraud database for years 2000-2012 and the Heritage Foundation's Election Fraud database for years 2013-2020. The News21 database was collected in 2012, while the Heritage Foundation began collecting in 2016. (The Heritage Foundation database does include fraud cases prior to 2013, although fewer than the News21 database.) We include all instances of fraud in the data—regardless of the charge, outcome, or whether the election was local, state, or federal—except we remove records with incomplete year or state information, yielding 2281 voter fraud cases between 2000-2020 across 50 states. All data were retrieved on October 1st, 2020. The number of cases per state is reported in Table 1 in the appendix.

The first three figures demonstrate the differential role mail-in ballots play in RBM states. Figure 1.a displays the percent of ballots cast that were received by mail in each state in the 2016 and 2018 general elections, as reported in the EAVS. The figure indicates that the percent of voters returning mailed ballots remained relatively unchanged between the 2016 and 2018 general elections—although, according to the 2018 EAVS report, "some states' data management systems do not distinguish in-person absentee voters from by-mail voters".[1] Figure 1.b compares the percent of ballots received by mail reported in the EAVS to the percent reported by the EP. The two datasets disagree on whether a majority of the ballots cast was received by mail in California, Hawaii, New Mexico, and Tennessee. Figure 1.c displays the percent of ballots received by mail that were rejected in each state in the 2016 and 2018 general elections, indicating that RBM states have among the lowest mail-in ballot rejection rates. (We could not find similar data on rejection rates from the EP.) These figures demonstrate meaningful differences between elections in RBM (VBM) and non-RBM (non-VBM); we argue that were fraud much more common when elections are held by mail, we would observe different fraud rates when comparing RBM (VBM) and non-RBM (non-VBM) states.

Figure 1.d suggests that states with more eligible voters tend to have more cases of reported fraud. It displays the log-number of fraud cases in the News21 dataset (top) and Heritage Foundation dataset (bottom) against the log-number of eligible voters in the most recent general election. States with zero fraud are censored. Best-fit lines are added for comparison (calculated using the Fisher scoring algorithm for a Tobit model with slope set to 1); for states on the line, a percent increase in the number of mail-in ballots corresponds with a percent increase in voter fraud cases. We argue fair comparisons between states should consider differences in the voting population, which may reflect the complexity of holding an election and the number of opportunities for fraud.

States such as Utah and Washington had zero fraud cases reported in the Heritage Foundation dataset after 2012. This suggests that voting by mail is relatively safe in Utah and Washington. However, it does not necessarily mean that vote-by-mail laws eliminated voter fraud in these states—or even that vote-by-mail laws reduced voter fraud, since non-VBM states may have experienced similar or larger declines. The following two sections compare RBM (VBM) with non-RBM (non-VBM) states to better evaluate whether RBM (VBM) states have increased their risk for voter fraud.

Section 2. Fraud rates are not higher in RBM (VBM) states than non-RBM (non-VBM) states

We compare the number of fraud cases per eligible voter in the most recent general election between two five-year periods: 2000-2004 and 2016-2020, which we refer to as the periods "following 2000" and "following 2016". We make this comparison for both RBM (VBM) states and non-RBM (non-VBM) states. We find that fraud rates were statistically indistinguishable between RBM (VBM) and non-RBM (non-VBM) states following 2016. We find no evidence that fraud is more prevalent among RBM or VBM states or that fraud has increased as vote-by-mail policies have expanded.

The five-year periods were chosen because all VBM states had VBM elections following 2016, while four out of five did not have VBM elections following 2000. The fifth state, Oregon, passed its VBM law in 1998, two years before any of our data were collected. However, the findings of this section hold when Oregon is omitted from the analysis.

Our results are displayed in Figures 2.a and 2.b in the appendix. In both plots, thick lines demarcate one standard error from the mean, while thin lines demarcate two standard errors. Figure 2.a indicates that RBM states changed slightly from an average of 0.49 fraud cases per million voters following 2000 to 0.31 following 2016. (The one-sided p-value is 0.16 using a t-distribution with four degrees of freedom.) In comparison, non-RBM states decreased in fraud from an average of 0.81 cases per million voters to 0.35 (p-value 0.05). RBM states had an average fraud rate 0.04 cases per million voters lower than non-RBM states (p-value 0.38).

Figure 2.b indicates that VBM states decreased from an average of 1.07 fraud cases per million voters following 2000 to 0.264 following 2016 (p-value 0.04). In comparison, non-VBM states decreased in fraud from an average of 0.73 cases per million voters to 0.35 (p-value 0.06). VBM states had an average fraud rate of 0.08 cases per million voters lower than non-VBM states (p-value 0.28).

Neither RBM nor VBM states have consistently higher fraud rates than non-RBM or non-VBM states. We note that between 2000 and 2016 the percent of ballots received by mail increased from one in ten to nearly one in four; if receiving ballots by mail increased fraud, one would expect RBM (VBM) states to have higher levels of fraud, not lower as we observe in the data.

One could argue from Figure 2.a that RBM states missed out on the 0.46 decline in fraud cases per million voters experienced by the non-RBM states. That is, had RBM states not had a majority of voters receive their ballots by mail, the fraud rate could have declined by 0.46 instead of the observed 0.18 decrease. This comparison of differences to differences is typically referred to, unsurprisingly, as a differences-in-differences analysis.[21] We disagree with this interpretation of the data since non-VBM states do not share a similar decline; in fact, VBM states declined by 0.81 while non-VBM declined only 0.38. We speculate that the smaller decline among RBM states is due, in part, to the fact that states with high fraud rates following 2000 were less likely to adopt RBM policies. To check this reasoning, we perform a second comparison of select VBM and non-VBM states in the following selection, which considers the possibility that fraud rates would have been higher had vote-by-mail states not adopted their vote-by-mail law.

Section 3. Washington and Colorado did not have higher fraud rates when compared to similar "synthetic" states that did not switch to voting by mail.

We estimate the number of fraud cases that would have occurred had Washington and Colorado not passed vote-by-mail laws. We find Washington would have reported eighty more cases of fraud had it not introduced its 2011 law (statistically different from zero) while Colorado would have reported one fewer case had it not introduced its 2013 law (not statistically different from zero). In neither case, do we find evidence that the number of fraud cases would have been higher had vote-by-mail laws not been passed.

We obtained these results using the matrix completion approach of Athey et al. (2018),[24] coded in the R package `gsynth`.[28,29] The approach excludes the fraud cases of vote-by-mail states in the years in which they had vote-by-mail elections. It then uses the remaining observations to estimate, via nuclear norm penalized

regression, the hypothetical amount of fraud that would have occurred in those years had elections not been vote-by-mail elections. The matrix completion approach builds on the synthetic controls literature[30–33] and "unconfoundedness" literature.[22,23,34] Following the synthetic controls literature, we refer to the hypothetical fraud a state would have experienced as "synthetic". For example, the fraud Washington would have reported had it not switched to VBM is called "synthetic Washington".

The matrix completion approach appears particularly robust to the large amount of zeros in the data; many states have years where zero cases of voter fraud are reported, see Table 1 in the appendix. We also control for the number of eligible voters in the most recent general election and the number of ballots for the highest office in the most recent general election. Unfortunately, there is not enough data to include additional information on mail in voting.

The results are displayed in figure Figures 3.a and 3.b. The solid line shows the annual number of reported fraud cases in Washington (Figures 3.a) and Colorado (Figures 3.b). The dotted line shows the synthetic alternative: the estimated number of cases that would have occurred had vote by mail not been adopted; the dark and light regions represent 50 percent and 95 percent bootstrapped intervals.

The vertical line shows the year the vote-by-mail law was adopted. Before the vertical line, synthetic Washington (Colorado) borrows data from states that have similar levels of baseline fraud, and thus the solid and dotted lines overlap. After the vertical line, we can observe the difference between the number of fraud cases observed for each state and the number that might have been observed had vote by mail not been adopted. In neither Washington nor Colorado do we find evidence that the number of fraud cases would have been higher had vote-by-mail laws not been passed. We only show comparisons for Washington and Colorado because Oregon became the first vote-by-mail state prior to our fraud data (in 1998) and Utah has no fraud cases except for 51 cases in 2006 connected to one election for sheriff of Daggett County.

Section 4. Discussion

We estimate the change in the reported number of voter fraud cases when states switch to conducting elections by mail. We believe our results contribute to the ongoing debate between proponents of election security and accessibility. We conclude by noting limitations of our analysis; we hope additional research will clarify and expand on the relationships reported here.

Reported voter fraud, like any reported crime, represents only a fraction of total cases. Our analysis tacitly assumes reported fraud is proportional to unreported fraud, and states and years with higher numbers of criminal investigations of voter fraud have higher amounts of voter fraud. However, our results could simply reflect non-idiosyncratic changes in how states conduct investigations or how News21/Heritage Foundation collected cases.

We include all reported voter fraud cases regardless of the evidence, charge, or outcome. We also do not distinguish between vote-by-mail laws and their implementation. It may be that the 2020 general election faces challenges not reflected by the 2016 and 2018 general elections. For example, we do not account for increased politicization or the capacity of the U.S. Postal Service. In addition, voters new to vote by mail may be different from voters that have voted by mail in the past; data suggest Democrats may be more likely to vote by mail. While there is certainly value in accounting for different types of fraud and VBM implementations, we believe the complex and subjective decisions necessary to account for these factors would undermine whatever conclusions might be reached from these analyses.

In a related point, Section 3 assumes vote by mail starts when state law requires that ballots are automatically sent to every registered voter. Prior to that law, large portions of each state may have already received ballots by mail. Select counties may automatically send ballots or a large number of voters may choose to request ballots by mail. For example, Washington "evolved" into a vote-by-mail state by first allowing voters to obtain permanent absentee status in 1993, and then allowing counties to opt into vote by mail in 2005. The estimates assume that, had vote-by-mail states not adopted vote by mail, older laws would still be in effect; it does not assume elections would have been entirely in person, which may be a relevant comparison for certain states new to vote by mail in the 2020 general election.

We rely on two different fraud datasets: from News21 (collected in 2012) and Heritage Foundation (collected beginning 2016). Since the two datasets have many cases in common, we decided to use the entire News21 dataset (2000-2012) and add additional years (2013-2020) from the Heritage data. As far as we can tell, this combination produces no artifacts: half of our findings only rely on the Heritage data, while the remaining findings essentially match states on the News21 data and evaluate RBM (VBM) policies with the Heritage data. Repeating the analysis with only the Heritage data does not contradict the evidence we present here.

A large amount of fraud comes from primaries and state and local elections. These elections typically have lower turnout, and local elections may exist only in select parts of the state. As a result, using the number of eligible voters in the most recent general election, as we do in this analysis, may understate the incidence of fraud. Many VBM states began voting by mail in local elections, and, unfortunately, we are unable to account for this in our analysis.

In summary, state-level comparisons are limited in what they can say about vote-by-mail policies. However, their simplicity and transparency make them indispensable tools for substantiating anecdotal evidence and expert opinion on election security and accessibility. In performing our comparisons, we find no evidence to suggest that voting by mail increases the risk of voter fraud overall. We believe our findings are unlikely were fraud much more common when elections are held by mail.

References

1. US Election Assistance Commission. *Election administration and voting survey.* https://www.eac.gov/research-and-data/studies-and-reports (2020).

2. National Conference of State Legislatures. *Voting outside the poling place.* www.ncsl.org/research/elections-and-campaigns/absentee-and-early-voting.aspx (2020).

3. Love, J., Stevens, M. & Gamio, L. Where americans can vote by mail in the 2020 elections. *The New York Times* (2020).

4. Glickhouse, R. Electionland 2020: Mail ballot challenges, election security, new legislation and more. *ProPublica* (2020).

5. Brennan Center for Justice. *Court case tracker: Voting rights litigation 2020. New York University School of Law* https://www.brennancenter.org/our-work/court-cases/voting-rights-litigation-2020 (2020).

6. project, S. healthy elections. *COVID-related election litigation tracker.* https://healthyelections-case-tracker.stanford.edu/ (2020).

7. Southwell, P. L. & Burchett, J. I. The effect of all-mail elections on voter turnout. *American Politics Quarterly* **28**, 72–79 (2000).

8. Berinsky, A. J., Burns, N. & Traugott, M. W. Who votes by mail?: A dynamic model of the individual-level consequences of voting-by-mail systems. *Public Opinion Quarterly* **65**, 178–197 (2001).

9. Alvarez, R. M., Hall, T. E. & Sinclair, B. Whose absentee votes are returned and counted: The variety and use of absentee ballots in california. *Electoral Studies* **27**, 673–683 (2008).

10. Smith, D. A. Vote-by-mail ballots cast in florida. *ACLU of Florida* (2018).

11. Hill, S. Election fraud references. (2006).

12. Brennan Center for Justice. *Resources on voter fraud claims.* https://www.brennancenter.org/our-work/research-reports/resources-voter-fraud-claims (2017).

13. Spakovsky, H. A. von. *Four stolen elections: The vulnerabilities of absentee and mail-in ballots.* https://www.heritage.org/election-integrity/report/four-stolen-elections-the-vulnerabilities-absentee-and-mail-ballots (2020).

14. Levitt, J. The truth about voter fraud. *Available at SSRN 1647224* (2007).

15. Mehrbani, R. Heritage fraud database: An assessment. *Brennan Center for Justice. New York University School of Law* (2017).

16. McReynolds, A. & Stewart III, C. Let's put the vote-by-mail 'fraud'myth to rest. *The Hill* (2020).

17. Viebeck, E. Minuscule number of potentially fraudulent ballots in states with universal mail voting undercuts trump claims about election risks. *The Washington Post* (2020).

18. Stark, P. B. Risk-limiting vote-tabulation audits: The importance of cluster size. *Chance* **23**, 9–12 (2010).

19. Lindeman, M. & Stark, P. B. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy* **10**, 42–49 (2012).

20. National Conference of State Legislatures. *Post-election audits*. www.ncsl.org/research/elections-and-campaigns/post-election-audits635926066.aspx (2019).

21. Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. (Princeton university press, 2008).

22. Rosenbaum, P. R. *Design of observational studies*. vol. 10 (Springer, 2010).

23. Imbens, G. W. & Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. (Cambridge University Press, 2015).

24. Athey, S., Bayati, M., Doudchenko, N., Imbens, G. & Khosravi, K. *Matrix completion methods for causal panel data models*. (2018).

25. McDonald, M. United states election project. (2020).

26. United States Election Assistance Commission. *EAVS deep dive: Early, absentee, and mail voting*. https://www.eac.gov/documents/2017/10/17/eavs-deep-dive-early-absentee-andmail-voting-data-statutory-overview (2017).

27. National Conference of State Legislatures. *State laws governing early voting*. https://www.ncsl.org/research/elections-and-campaigns/early-voting-in-state-elections.aspx (2020).

28. Xu, Y. & Liu, L. *Gsynth: Generalized synthetic control method*. (2018).

29. Xu, Y. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* **25**, 57–76 (2017).

30. Abadie, A. & Gardeazabal, J. The economic costs of conflict: A case study of the basque country. *American economic review* **93**, 113–132 (2003).

31. Abadie, A., Diamond, A. & Hainmueller, J. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association* **105**, 493–505 (2010).

32. Ben-Michael, E., Feller, A. & Rothstein, J. New perspectives on the synthetic control method. (2018).

33. Ben-Michael, E., Feller, A. & Rothstein, J. The augmented synthetic control method. (2020).

34. Rosenbaum, P. R. Observational studies. (2002).

Appendix

Figure 1.a. Roughly the same percentage of ballots were received by mail in the 2016 and 2018 general elections according to EAVS

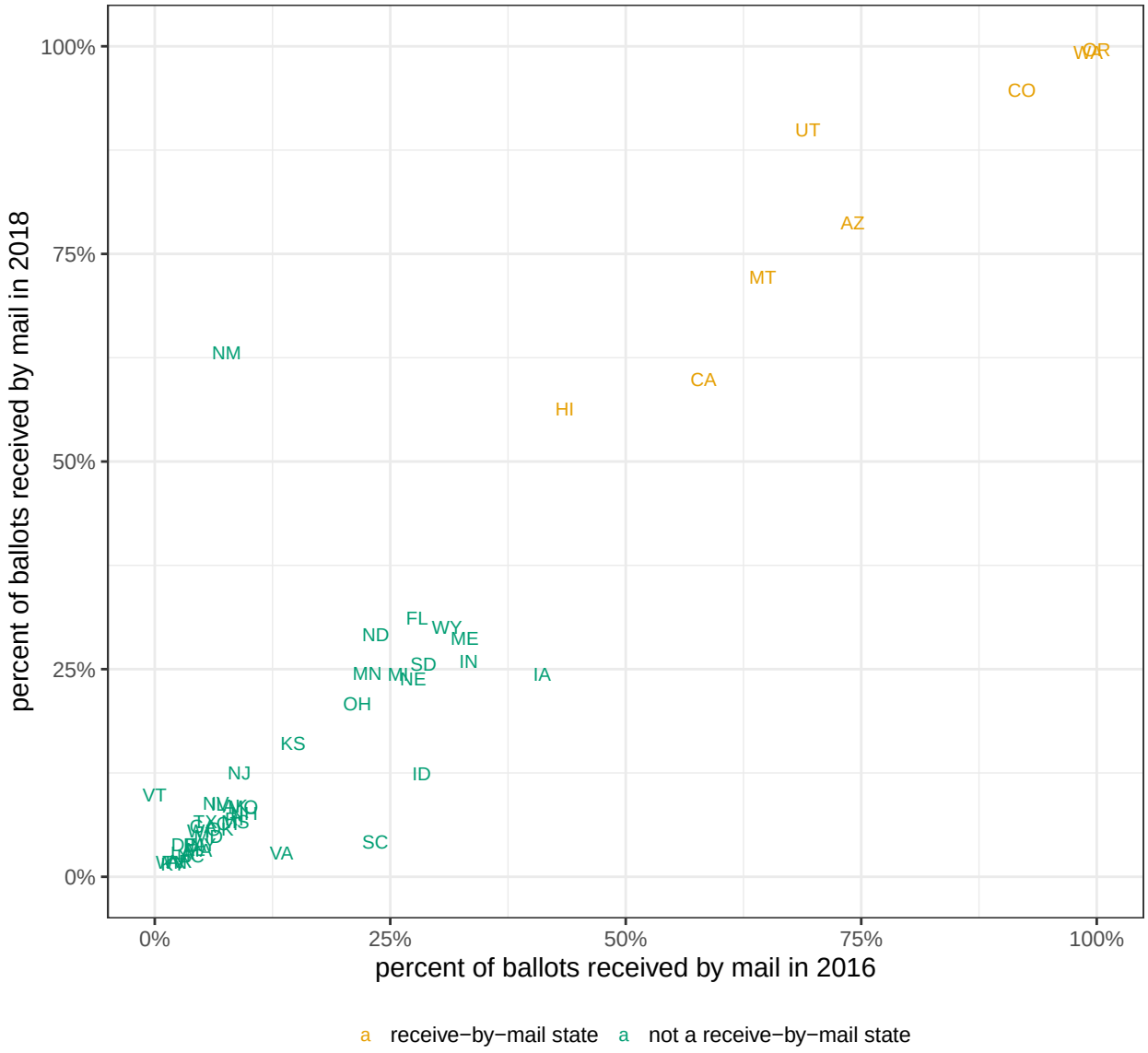

Figure 1.b. EAVS and EP report different percentage of ballots were received by mail in the 2018 general election

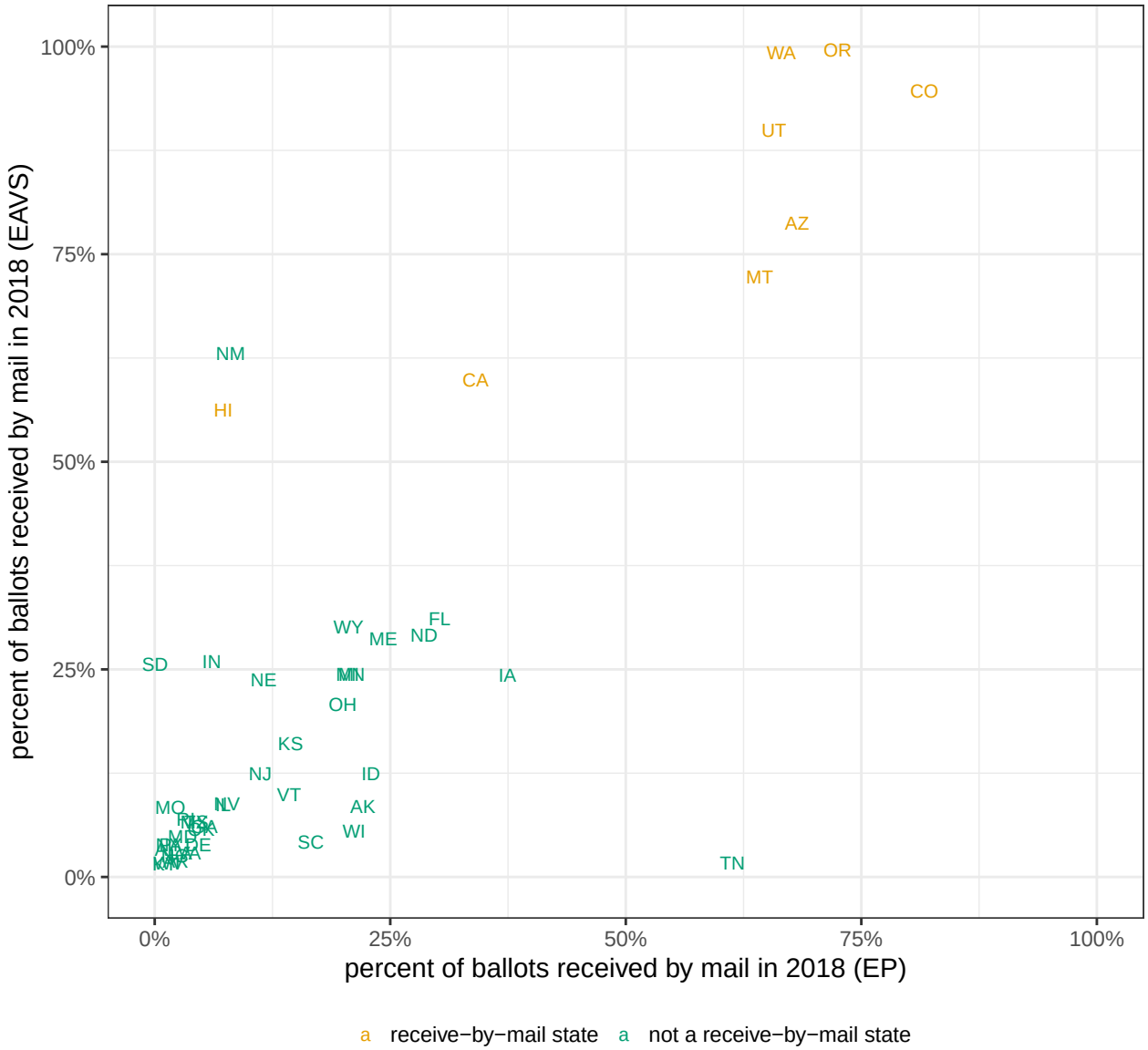

Figure 1.c. Receive−by−mail states had a lower percentage of mail in ballots rejected in the 2016 and 2018 general elections

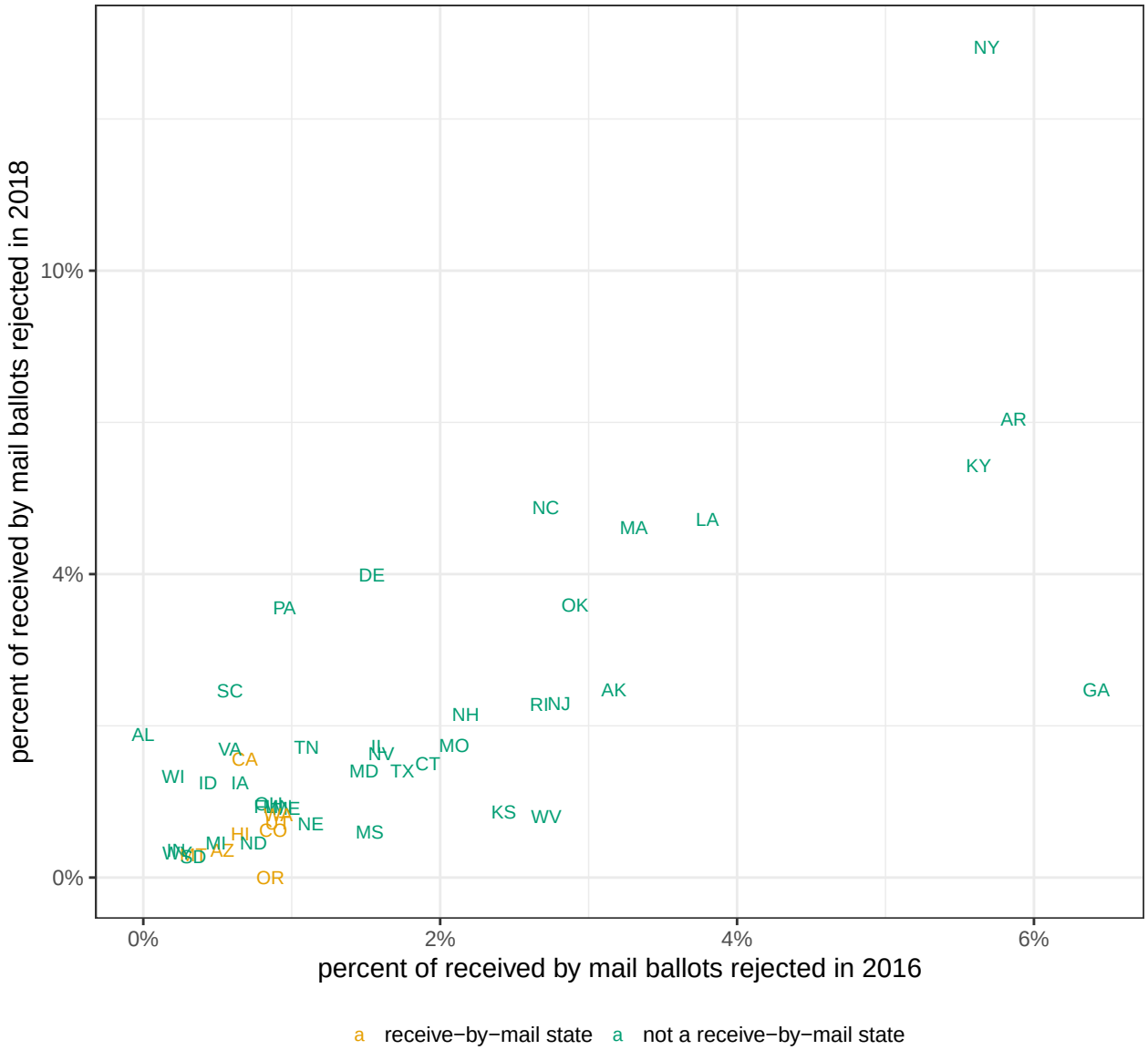

Figure 1.d. States with more eligible voters have more cases of reported fraud, although there is considerable variation across states in both the News21 and Heritage Foundation databases.

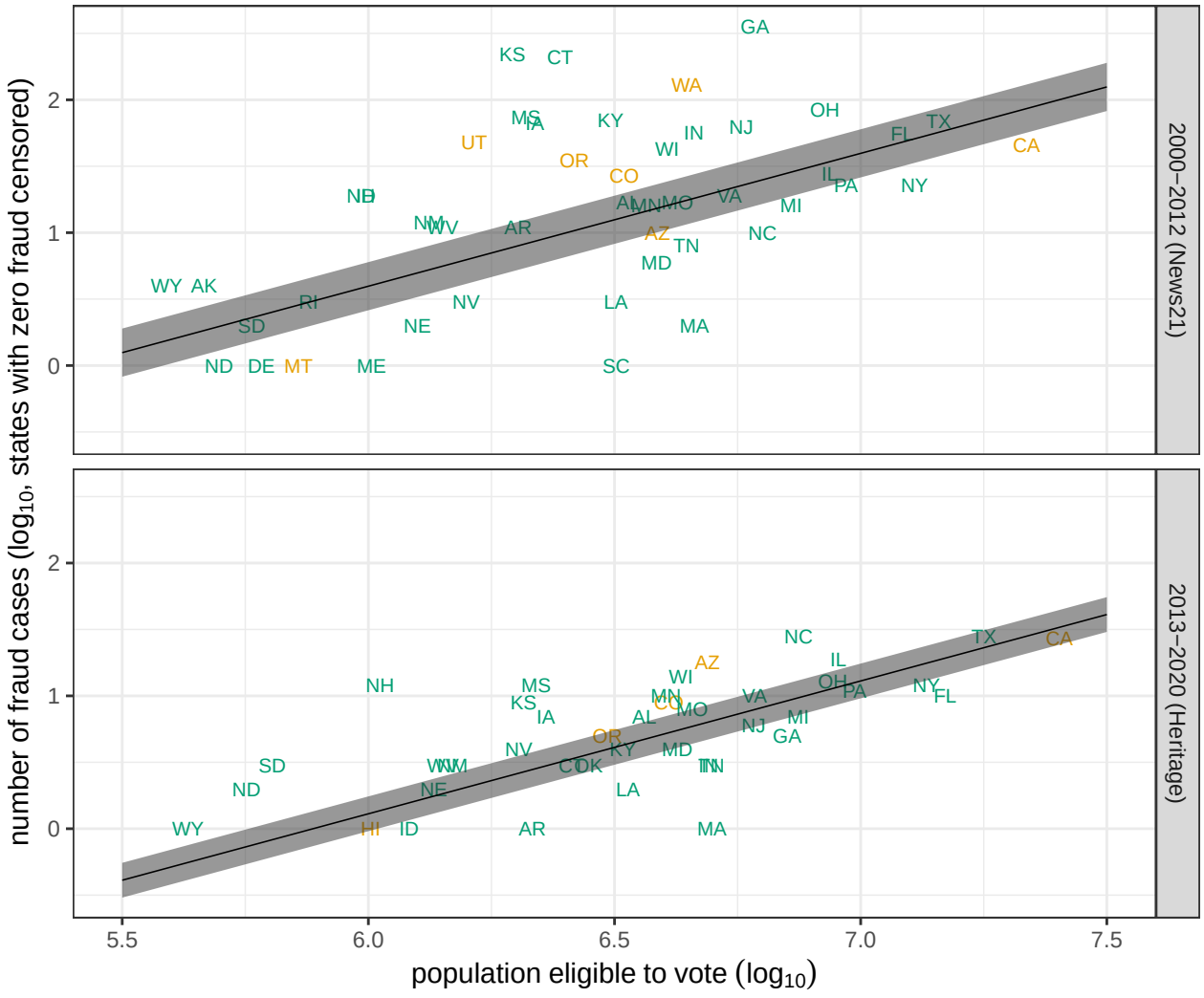

Figure 2.a. Fraud per voter declined thirty−six percent in states where a majority receive by mail and fifty−six percent in other states

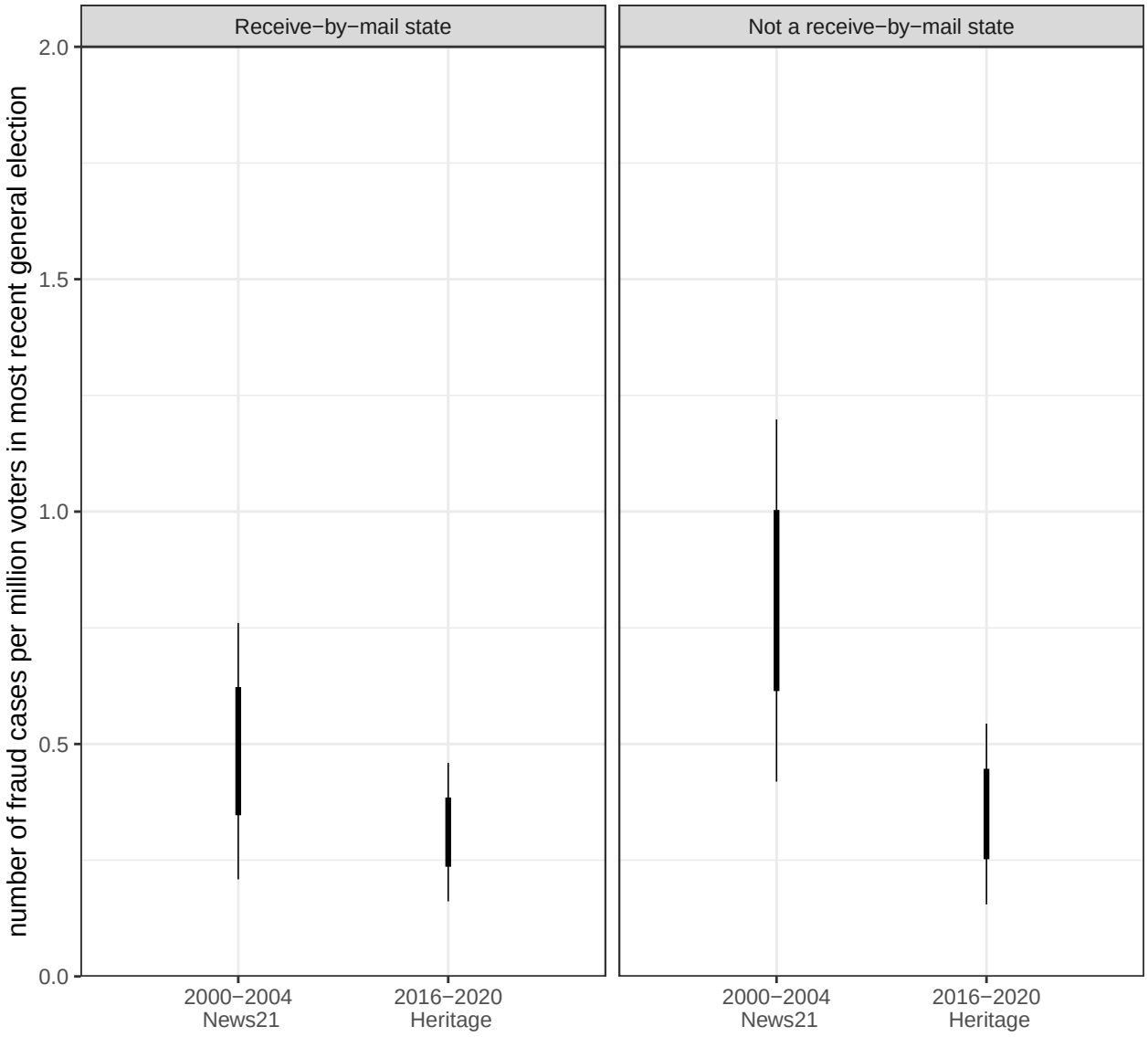

Figure 2.b. Fraud per voter declined seventy−five percent in vote−by−mail states and fifty−two percent in other states

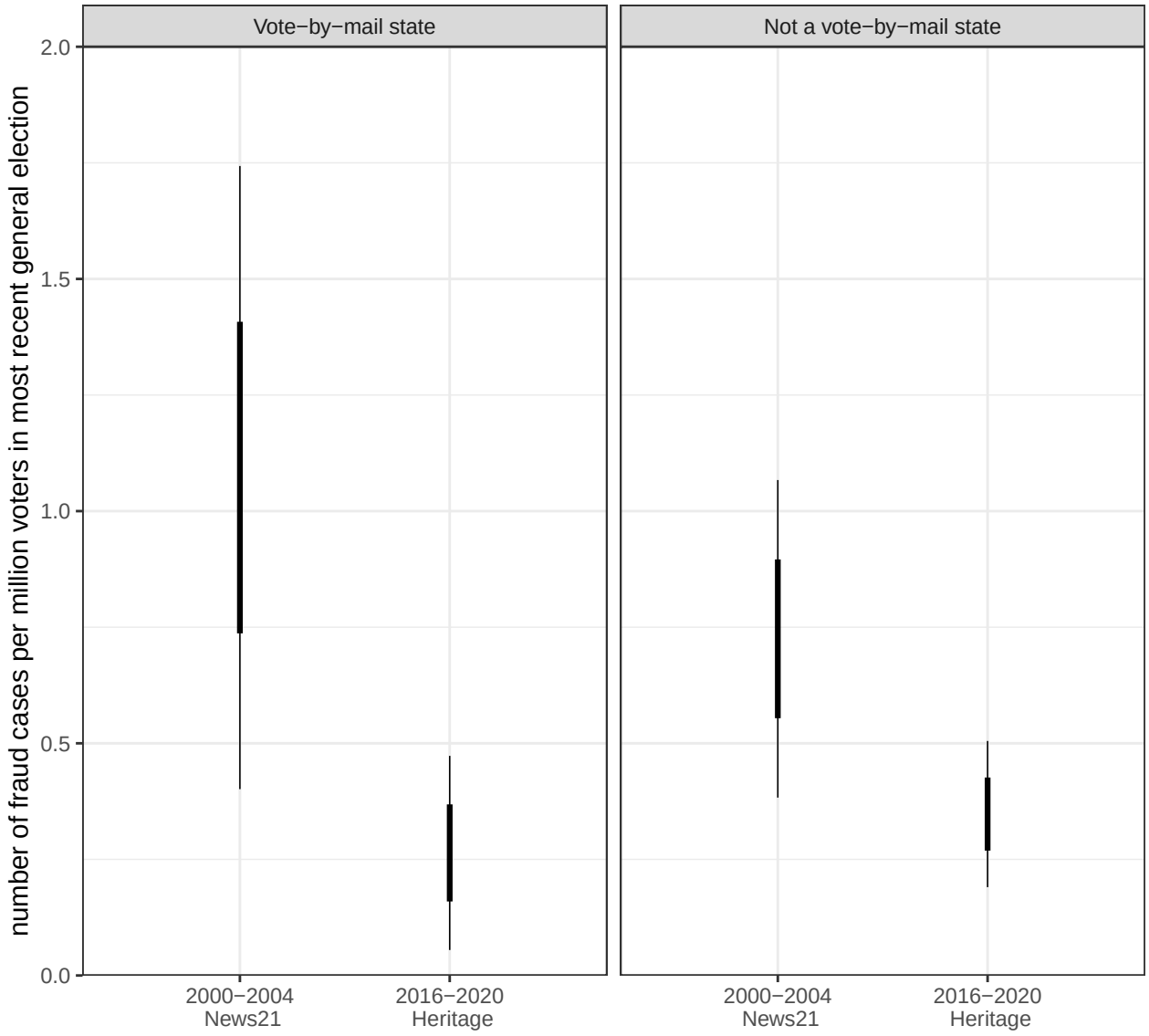

Figure 3.a. Washington would have eighty more fraud cases had it not instituted its vote−by−mail policy (± ten cases).

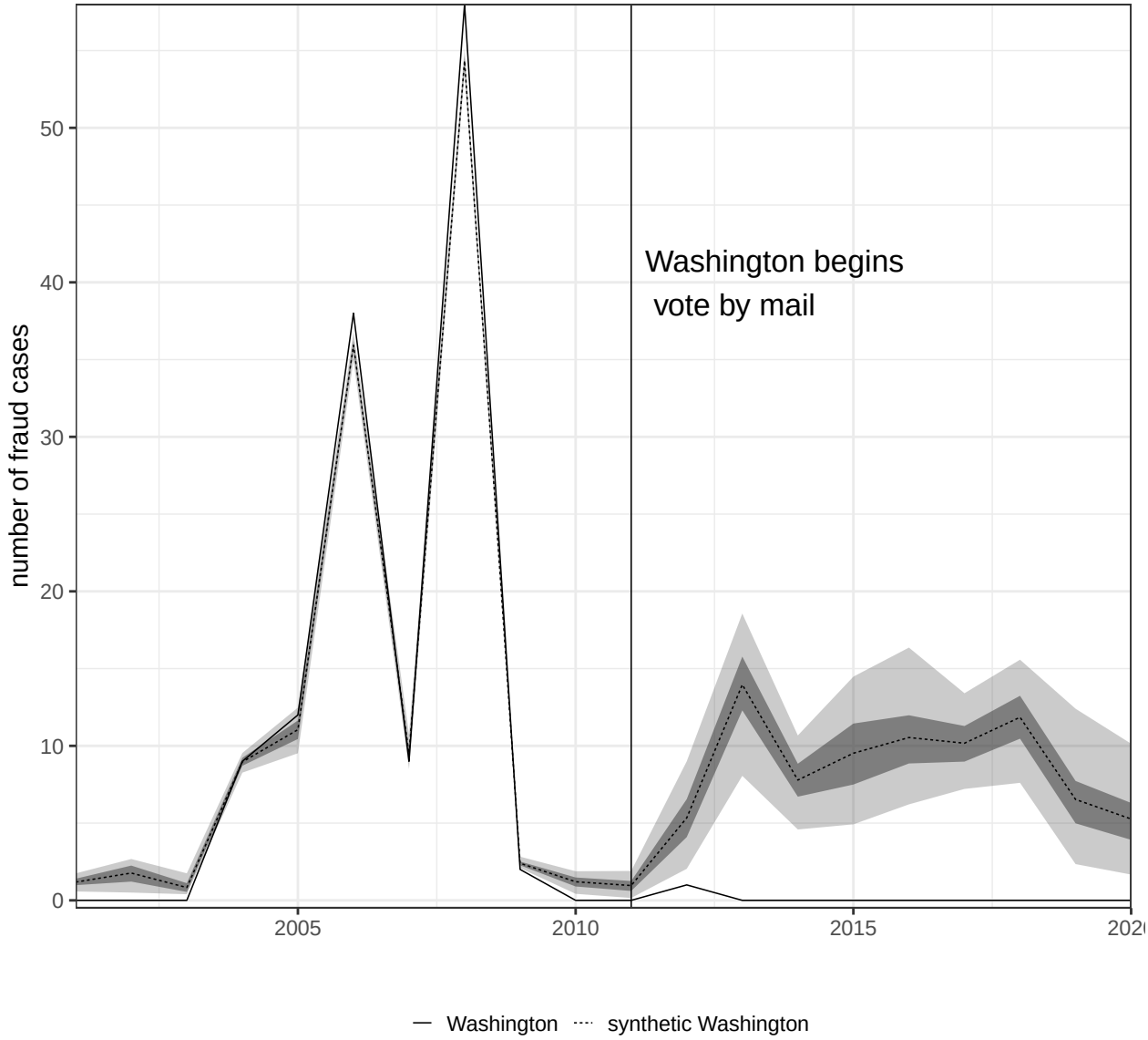

Figure 3.b. Colorado would have one less fraud case had it not instituted its vote−by−mail policy (± two cases).

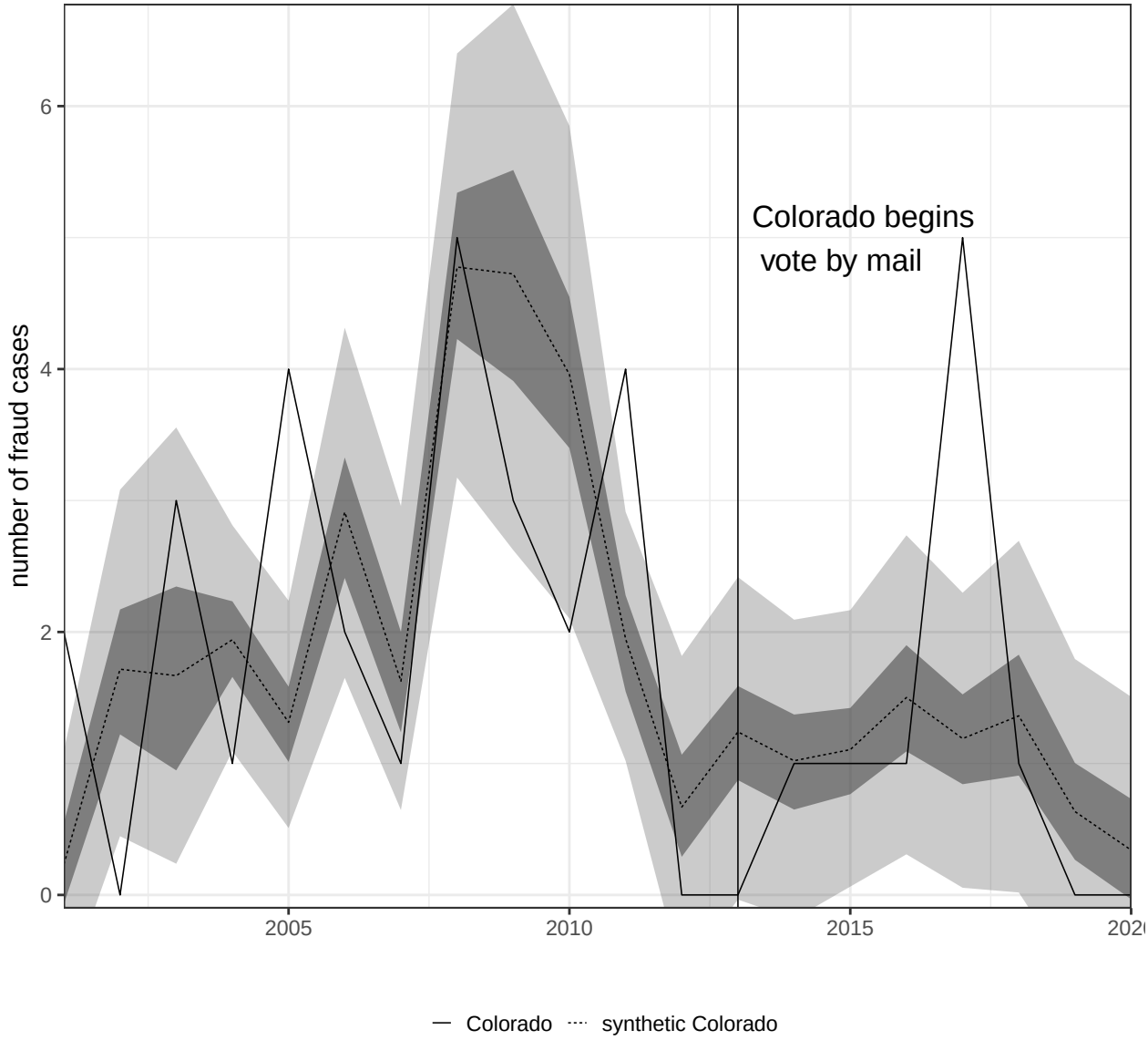

Table 1: The number of reported fraud cases by state and year. Incomplete/redundant records removed.

State	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Alabama	10	1	1	0	1	0	0	0	2	0	2
Alaska	0	0	0	0	0	1	0	0	0	0	0
Arizona	0	0	0	0	0	2	0	0	3	0	0
Arkansas	0	0	2	0	0	0	1	0	0	4	2
California	2	2	2	1	4	6	19	0	1	2	4
Colorado	0	2	0	3	1	4	2	1	5	3	2
Connecticut	4	5	34	17	22	14	11	9	21	33	10
Delaware	0	0	0	0	0	0	1	0	0	0	0
Florida	1	0	0	1	14	1	1	0	3	5	25
Georgia	0	0	1	1	6	13	29	29	33	79	109
Hawaii	0	0	0	0	0	0	0	0	0	0	0
Idaho	0	0	0	0	3	0	0	0	1	5	3
Illinois	0	0	1	1	10	1	1	0	7	3	0
Indiana	0	0	0	43	0	0	0	1	4	0	0
Iowa	0	0	0	0	5	5	2	2	2	19	23
Kansas	15	0	50	0	10	8	17	0	61	7	50
Kentucky	5	2	5	8	7	1	11	1	11	7	13
Louisiana	0	0	2	1	0	0	0	0	0	0	0
Maine	0	0	0	0	0	0	0	0	1	0	0
Maryland	0	0	0	0	0	0	0	0	0	0	3
Massachusetts	0	0	0	0	0	0	0	0	0	0	0
Michigan	0	4	3	0	1	1	1	0	0	0	5
Minnesota	0	0	0	0	0	0	0	0	12	1	1
Mississippi	1	4	1	1	2	1	1	23	6	20	6
Missouri	0	2	0	0	1	1	13	0	0	0	0
Montana	0	0	0	0	0	0	0	0	0	0	0
Nebraska	0	0	0	0	1	0	0	0	0	0	1
Nevada	0	0	0	0	0	0	0	0	2	0	0
New Hampshire	0	0	0	0	2	0	3	0	10	0	3
New Jersey	0	0	0	0	0	0	1	12	0	45	4
New Mexico	0	0	0	0	1	0	0	0	0	0	0
New York	0	0	0	0	0	4	0	1	1	15	0
North Carolina	1	0	4	0	0	0	0	0	4	0	0
North Dakota	0	0	0	0	0	1	0	0	0	0	0
Ohio	2	1	0	0	4	2	19	12	28	7	0
Oklahoma	0	0	0	0	0	0	0	0	0	0	0
Oregon	7	4	2	3	6	4	4	1	0	3	0
Pennsylvania	0	0	1	1	1	0	0	0	2	17	0
Rhode Island	0	0	0	0	2	0	0	1	0	0	0
South Carolina	0	0	0	0	0	0	0	1	0	0	0
South Dakota	0	0	0	1	0	0	0	0	0	0	0
Tennessee	0	0	0	0	0	1	1	1	0	0	0
Texas	0	0	1	1	3	0	20	5	24	4	1
Utah	0	0	0	0	0	0	48	0	0	0	0
Vermont	0	0	0	0	0	0	0	0	0	0	0
Virginia	1	2	1	1	0	2	3	0	4	0	1
Washington	0	0	0	0	9	12	38	9	58	2	0

(continued)

State	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
West Virginia	0	0	0	0	1	0	9	0	0	0	0
Wisconsin	0	0	0	0	3	14	0	0	13	0	1
Wyoming	1	0	0	0	0	0	1	0	1	0	0

State	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Alabama	2	0	0	0	0	3	1	2	0	1	0
Alaska	0	0	2	1	0	0	0	0	0	0	0
Arizona	0	5	0	0	6	3	7	0	0	1	1
Arkansas	2	2	0	0	0	0	1	0	0	0	0
California	4	2	1	0	4	2	6	3	1	3	8
Colorado	2	4	0	0	1	1	1	5	1	0	0
Connecticut	10	28	0	1	0	1	0	0	2	0	0
Delaware	0	0	0	0	0	0	0	0	0	0	0
Florida	25	0	0	5	1	3	0	5	1	0	0
Georgia	109	57	2	0	0	1	0	3	1	0	0
Hawaii	0	0	0	0	0	0	1	0	0	0	0
Idaho	3	7	0	0	0	0	0	1	0	0	0
Illinois	0	2	0	2	1	1	2	3	12	0	0
Indiana	0	4	0	5	0	0	1	2	0	0	0
Iowa	23	4	2	3	4	0	2	1	0	0	0
Kansas	50	0	2	0	0	1	6	2	0	0	0
Kentucky	13	0	0	0	1	0	2	1	0	0	0
Louisiana	0	0	0	0	0	1	0	0	0	0	1
Maine	0	0	0	0	0	0	0	0	0	0	0
Maryland	3	0	0	3	1	0	0	1	2	0	0
Massachusetts	0	0	0	2	0	1	0	0	0	0	0
Michigan	5	0	0	1	5	1	1	0	0	0	0
Minnesota	1	0	0	2	0	1	6	1	2	0	0
Mississippi	6	1	6	1	2	3	0	0	1	6	0
Missouri	0	0	0	0	0	1	4	3	0	0	0
Montana	0	1	0	0	0	0	0	0	0	0	0
Nebraska	1	0	0	0	0	0	0	2	0	0	0
Nevada	0	0	0	1	2	0	1	1	0	0	0
New Hampshire	3	0	0	1	1	0	2	1	3	5	0
New Jersey	4	0	1	0	3	1	0	1	0	1	0
New Mexico	0	0	10	1	0	1	0	0	0	0	2
New York	0	0	1	1	2	1	4	3	1	1	0
North Carolina	0	1	0	0	2	2	5	4	13	2	0
North Dakota	0	0	0	0	0	0	0	1	1	0	0
Ohio	0	0	0	9	2	1	0	3	5	2	0
Oklahoma	0	0	0	0	0	0	0	0	1	2	0
Oregon	0	0	0	1	0	0	0	0	0	5	0
Pennsylvania	0	1	0	0	1	1	3	1	4	1	0
Rhode Island	0	0	0	0	0	0	0	0	0	0	0
South Carolina	0	0	0	0	0	0	0	0	0	0	0
South Dakota	0	0	0	1	0	3	0	0	0	0	0
Tennessee	0	0	5	0	1	0	0	0	1	1	0

(continued)

State	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Texas	1	4	0	6	4	7	2	6	8	1	0
Utah	0	0	0	0	0	0	0	0	0	0	0
Vermont	0	0	0	0	0	0	0	0	0	0	0
Virginia	1	1	0	3	2	0	2	2	3	0	1
Washington	0	0	1	0	0	0	0	0	0	0	0
West Virginia	0	0	0	1	0	0	0	1	1	0	1
Wisconsin	1	0	2	10	4	3	2	4	1	0	0
Wyoming	0	0	1	0	1	0	0	0	0	0	0

```
##################
# Voting by Mail #
##################

# Date: 10/20/2020

# This R script investigates whether election fraud is higher among U.S. states
## with vote-by-mail laws or where many ballots are received by mail.
## Section 1. examines which states are vote-by-mail/receive-by-mail states.
## Section 2. examines fraud on aggregate across these states. Section 3.
## estimates what fraud would have looked like if Washington and Colorado had not
## passed vote-by-mail laws, using the gsynth package. Fraud data is from the
## Heritage Foundation voter fraud database (https://www.heritage.org/voterfraud)
## and News21 election fraud database
## (https://votingrights.news21.com/interactive/election-fraud-database/). Data
## on elections is from the U.S. Election Assistance Commission Election
## Administration and Voting Surveys
## (https://www.eac.gov/research-and-data/datasets-codebooks-and-surveys) and the
## United States Election Project
## (http://www.electproject.org/home/voter-turnout/voter-turnout-data)

# Please contact Jonathan Auerbach with any questions or corrections: jonathan@amstat.org
library("tidyverse")
library("VGAM")

# Heritage Foundation voter fraud database
## source: heritage.org/voterfraud
heritage <- read_csv("hfd.csv")
# News21 election fraud database
## source: votingrights.news21.com/interactive/election-fraud-database/
news21 <- read_csv("news21.csv")
# U.S. Election Assistance Commission Election Administration and Voting Surveys
## source: eac.gov/research-and-data/datasets-codebooks-and-surveys
### Overview Table 2:
### eac.gov/sites/default/files/eac_assets/1/6/2018_EAVS_Report.pdf
eavs_2018 <- read_csv("eavs_2018_table_2.csv") %>%
  filter(State %in% datasets::state.name)
### Overview Table 2:
### eac.gov/sites/default/files/eac_assets/1/6/2016_EAVS_Comprehensive_Report.pdf
eavs_2016 <- read_csv("eavs_2016_table_2.csv") %>%
  filter(State %in% datasets::state.name)
# United States Election Project
## source: electproject.org/home/voter-turnout/voter-turnout-data
voting_eligible_population <- read_csv("voting_eligible_population.csv")
ballots_highest_office <-  read_csv("ballots_highest_office.csv")
ballots_highest_office$`2020` <- ballots_highest_office$`2018`

#list of vote-by-mail states
VBM <- c("Colorado", "Hawaii", "Oregon", "Utah", "Washington")
#list of receive-by-mail states (states where > 50% vote through mail)
RBM <- c("Arizona", "California", "Colorado", "Hawaii",
         "Montana", "Oregon", "Utah", "Washington")
```

```r
#Section 1
eavs_2018 %>%
  select(`Total By-Mail Ballots Returned 2018` = `Total By-Mail Ballots Returned`,
         `Total Voter Turnout 2018` = `Total Voter Turnout`,
         State) %>%
  left_join(eavs_2016 %>%
              select(`Total By-Mail Ballots Returned 2016` =
                       `Total By-Mail Ballots Returned`,
                     `Total Voter Turnout 2016` = `Total Voter Turnout`,
                     State)) %>%
  left_join(data.frame(State.abb = datasets::state.abb,
                       State = datasets::state.name)) %>%
  ggplot() +
  aes(`Total By-Mail Ballots Returned 2016`/`Total Voter Turnout 2016`,
      `Total By-Mail Ballots Returned 2018`/`Total Voter Turnout 2018`,
      color = relevel(factor(ifelse(State %in% RBM,
                                    "receive-by-mail state",
                                    "not a receive-by-mail state")),
                      "receive-by-mail state")) +
  geom_text(aes(label = State.abb), size = 5) +
  labs(color = "",
       y = "percent of ballots received by mail in 2018",
       x = "percent of ballots received by mail in 2016",
       title =
         "Figure 1.a. Roughly the same percentage of ballots were received by mail
                  in the 2016 and 2018 general elections according to EAVS") +
  theme(legend.position = "bottom") +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1),
                     limits = c(0, 1)) +
  scale_x_continuous(labels = scales::percent_format(accuracy = 1),
                     limits = c(0, 1)) +
  scale_colour_manual(values = c("#E69F00", "#009E73"))

eavs_2018 %>%
  select(`Total By-Mail Ballots Returned (EAVS)` =
           `Total By-Mail Ballots Returned`,
         `Total Voter Turnout (EAVS)` = `Total Voter Turnout`,
         State) %>%
  left_join(ep_2018 %>%
              select(`Total By-Mail Ballots Returned (EP)` =
                       `Mail/Online Ballots Returned`,
                     `Total Voter Turnout (EP)` = `2018 Vote for Highest Office`,
                     State)) %>%
  left_join(data.frame(State.abb = datasets::state.abb,
                       State = datasets::state.name)) %>%
  ggplot() +
  aes(`Total By-Mail Ballots Returned (EP)`/`Total Voter Turnout (EP)`,
      `Total By-Mail Ballots Returned (EAVS)`/`Total Voter Turnout (EAVS)`,
      color = relevel(factor(ifelse(State %in% RBM,
                                    "receive-by-mail state",
                                    "not a receive-by-mail state")),
                      "receive-by-mail state")) +
  geom_text(aes(label = State.abb), size = 5) +
```

```
  labs(color = "",
       y = "percent of ballots received by mail in 2018 (EAVS)",
       x = "percent of ballots received by mail in 2018 (EP)",
       title =
         "Figure 1.b. EAVS and EP report different percentage of ballots were
                    received by mail in the 2018 general election") +
  theme(legend.position = "bottom") +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1),
                     limits = c(0, 1)) +
  scale_x_continuous(labels = scales::percent_format(accuracy = 1),
                     limits = c(0, 1)) +
  scale_colour_manual(values = c("#E69F00", "#009E73"))

eavs_2018 %>%
  select(`Total By-Mail Ballots Returned 2018` = `Total By-Mail Ballots Returned`,
         `Total By-Mail Ballots Rejected 2018` = `Total By-Mail Ballots Rejected`,
         `Total Voter Turnout 2018` = `Total Voter Turnout`,
         State) %>%
  left_join(eavs_2016 %>%
              select(`Total By-Mail Ballots Returned 2016` =
                       `Total By-Mail Ballots Returned`,
                     `Total By-Mail Ballots Rejected 2016` =
                       `Total By-Mail Ballots Rejected`,
                     State)) %>%
  left_join(data.frame(State.abb = datasets::state.abb,
                       State = datasets::state.name)) %>%
  filter(State != "New Mexico") %>%
  ggplot() +
  aes(`Total By-Mail Ballots Rejected 2016`/`Total By-Mail Ballots Returned 2016`,
      `Total By-Mail Ballots Rejected 2018`/`Total By-Mail Ballots Returned 2018`,
      color = relevel(factor(ifelse(State %in% RBM,
                                    "receive-by-mail state",
                                    "not a receive-by-mail state")),
                      "receive-by-mail state")) +
  geom_text(aes(label = State.abb), size = 5) +
  labs(color = "",
       y = "percent of received by mail ballots rejected in 2018",
       x = "percent of received by mail ballots rejected in 2016",
       title =
         "Figure 1.c. Receive-by-mail states had a lower percentage of mail in
         ballots rejected in the 2016 and 2018 general elections") +
  theme(legend.position = "bottom") +
  scale_y_continuous(labels = scales::percent_format(accuracy = 2)) +
  scale_x_continuous(labels = scales::percent_format(accuracy = 2)) +
  scale_colour_manual(values = c("#E69F00", "#009E73"))

regression_data <-
  voting_eligible_population %>%
  gather(year, `Voting eligible population`, -state) %>%
  mutate(year = as.numeric(year)) %>%
  left_join(heritage %>%
              filter(YEAR > 2012) %>%
              select(STATE, YEAR) %>%
```

```
              mutate(YEAR = ifelse(YEAR %% 2 == 0, YEAR, YEAR - 1)) %>%
              rbind(news21 %>%
                      rename("STATE ABBREV" = "STATE") %>%
                      left_join(tibble(STATE = datasets::state.name,
                                       "STATE ABBREV" = datasets::state.abb)) %>%
                      select(STATE, YEAR) %>%
                      mutate(YEAR = ifelse(YEAR %% 2 == 0, YEAR, YEAR - 1))) %>%
              group_by(state = STATE, year = YEAR) %>%
              summarize(count = n()))  %>%
  mutate(count = ifelse(is.na(count), 0, count),
         Period = factor(ifelse(year <= 2012,
                                "2000-2012 (News21)",
                                "2013-2020 (Heritage)"),
                         levels = c("2000-2012 (News21)",
                                    "2013-2020 (Heritage)"))) %>%
  group_by(state, Period) %>%
  summarize(count = sum(count),
            `Voting eligible population` = mean(`Voting eligible population`)) %>%
  filter(state %in% datasets::state.name) %>%
  left_join(tibble(state = datasets::state.name,
                   `state abbrev` = datasets::state.abb))

tobit_before <-
  vglm(log10(count) ~ offset(log10(`Voting eligible population`)),
       tobit(Lower = 0),
       data = regression_data,
       subset = Period == "2000-2012 (News21)") %>%
  summary()

tobit_after <-
  vglm(log10(count) ~ offset(log10(`Voting eligible population`)),
       tobit(Lower = 0),
       data = regression_data,
       subset = Period == "2013-2020 (Heritage)") %>%
  summary()

regression_coefficients <-
  tibble(Period = factor(c(rep("2000-2012 (News21)", 2),
                           rep("2013-2020 (Heritage)", 2)),
                         levels = c("2000-2012 (News21)",
                                    "2013-2020 (Heritage)")),
         estimate = c(tobit_before@coef3["(Intercept):1",c("Estimate",
                                                           "Std. Error")],
                      tobit_after@coef3["(Intercept):1",c("Estimate",
                                                          "Std. Error")]),
         type = c("Estimate", "Std. Error","Estimate", "Std. Error")) %>%
  spread(type, estimate)

tobit_fit <-
  tibble(x = rep(5:7 + .5, 2),
         Period = factor(c(rep("2000-2012 (News21)", 3),
                           rep("2013-2020 (Heritage)", 3)),
                         levels = c("2000-2012 (News21)",
```

```
                                    "2013-2020 (Heritage)")),
         y = c(regression_coefficients$Estimate[1] + 5:7 + .5,
               regression_coefficients$Estimate[2] + 5:7 + .5),
         ymin = c(regression_coefficients$Estimate[1] + 5:7 + .5 -
                    2 * regression_coefficients$`Std. Error`[1],
                  regression_coefficients$Estimate[2] + 5:7 + .5 -
                    2 * regression_coefficients$`Std. Error`[2]),
         ymax = c(regression_coefficients$Estimate[1] + 5:7 + .5 +
                    2 * regression_coefficients$`Std. Error`[1],
                  regression_coefficients$Estimate[2] + 5:7 + .5 +
                    2 * regression_coefficients$`Std. Error`[2]))

regression_data %>%
  filter(count > 0) %>%
  ggplot() +
  geom_text(aes(log10(`Voting eligible population`), log10(count),
                label = `state abbrev`,
                color = relevel(factor(ifelse(state %in% RBM,
                            "receive-by-mail state before 2016",
                            "not a receive-by-mail state before 2016")),
                            "receive-by-mail state before 2016")),
            size = 5) +
  geom_ribbon(aes(x, ymin = ymin, ymax = ymax), data = tobit_fit, alpha = .5) +
  geom_line(aes(x, y), data = tobit_fit) +
  facet_grid(Period ~ .) +
  theme(legend.position = "bottom") +
  labs(x = expression(population~eligible~to~vote~(log[10])),
       y = expression(number~of~fraud~cases~"("*log[10]*",
                      "~states~with~zero~fraud~censored*")"),
       color = "",
       title = "Figure 1.d. States with more eligible voters have more cases of
       reported fraud, although there is considerable variation across states in
       both the News21 and Heritage Foundation databases.") +
  scale_colour_manual(values = c("#E69F00", "#009E73"))

#Section 2
#receive by mail
eavs_2016 <- read_csv("eavs_2016_table_2.csv") %>%
  filter(State %in% datasets::state.name) %>%
  select(State, `Total Voter Turnout 2016` = `Total Voter Turnout`)
eavs_2000 <- read_csv("eavs_2000.csv") %>%
  filter(STATE %in% datasets::state.name) %>%
  select(State = STATE, `Total Voter Turnout 2000` = TURNOUT)

year_state <-
  expand.grid(YEAR = c(2000:2004, 2016:2020),
              STATE = datasets::state.name) %>%
  mutate(`VBM Status` = ifelse(STATE %in% RBM,
                               "Receive-by-mail state",
                               "Not a receive-by-mail state"))
combined_data <-
  heritage %>%
  filter(YEAR %in% 2016:2020) %>%
```

```
  left_join(eavs_2016, by = c("STATE" = "State")) %>%
  group_by(YEAR, STATE) %>%
  summarize(count = n()) %>%
  rbind(news21 %>%
          filter(YEAR %in% 2000:2004) %>%
          rename("STATE ABBREV" = "STATE") %>%
          left_join(tibble(STATE = datasets::state.name,
                           "STATE ABBREV" = datasets::state.abb)) %>%
          group_by(YEAR, STATE) %>%
          summarize(count = n()))

fraud <-
  year_state %>%
  left_join(combined_data) %>%
  left_join(eavs_2016, by = c("STATE" = "State")) %>%
  left_join(eavs_2000, by = c("STATE" = "State")) %>%
  mutate(count = ifelse(is.na(count), 0, count),
         pop   = ifelse(YEAR < 2010,
                        `Total Voter Turnout 2000`,
                        `Total Voter Turnout 2016`)) %>%
  group_by(YEAR, `VBM Status`) %>%
  summarize(count = sum(count),
            pop   = sum(pop)) %>%
  mutate(Period = factor(ifelse(YEAR < 2010,
                                "2000-2004\n News21",
                                "2016-2020\n Heritage"),
                         levels = c("2000-2004\n News21",
                                    "2016-2020\n Heritage")))

fraud %>%
  group_by(`VBM Status`, Period) %>%
  summarize(n = n(),
            count_mean = mean(1e6 * count/pop),
            count_sd   = sd(1e6 * count/pop)/sqrt(n)) %>%
  ggplot() +
  aes(x = Period,
      y = count_mean) +
  geom_linerange(aes(ymin = count_mean - count_sd,
                     ymax = count_mean + count_sd),
                 size = 2) +
  geom_linerange(aes(ymin = count_mean - 2 * count_sd,
                     ymax = count_mean + 2 * count_sd)) +
  labs(x = "",
       y = "number of fraud cases per million voters in most recent general election",
       color = "",
       title = "Figure 2.a. Fraud per voter remained largely unchanged in states
       where a majority receive by mail, declined fifty-six percent elsewhere") +
  facet_wrap(~ relevel(factor(`VBM Status`), "Receive-by-mail state")) +
  coord_cartesian(ylim = c(0, 2), xlim = c(.25, 2.75), expand = FALSE)

#vote by mail
year_state <-
  expand.grid(YEAR = c(2000:2004, 2016:2020),
```

```
              STATE = datasets::state.name) %>%
  mutate(`VBM Status` = ifelse(STATE %in% VBM,
                               "Vote-by-mail state",
                               "Not a vote-by-mail state"))
combined_data <-
  heritage %>%
  filter(YEAR %in% 2016:2020) %>%
  left_join(eavs_2016, by = c("STATE" = "State")) %>%
  group_by(YEAR, STATE) %>%
  summarize(count = n()) %>%
  rbind(news21 %>%
          filter(YEAR %in% 2000:2004) %>%
          rename("STATE ABBREV" = "STATE") %>%
          left_join(tibble(STATE = datasets::state.name,
                           "STATE ABBREV" = datasets::state.abb)) %>%
          group_by(YEAR, STATE) %>%
          summarize(count = n()))

fraud <- year_state %>%
  left_join(combined_data) %>%
  left_join(eavs_2016, by = c("STATE" = "State")) %>%
  left_join(eavs_2000, by = c("STATE" = "State")) %>%
  mutate(count = ifelse(is.na(count), 0, count),
         pop   = ifelse(YEAR < 2010, `Total Voter Turnout 2000`,
                        `Total Voter Turnout 2016`)) %>%
  group_by(YEAR, `VBM Status`) %>%
  summarize(count = sum(count),
            pop   = sum(pop)) %>%
  mutate(Period = factor(ifelse(YEAR < 2010, "2000-2004\n News21",
                                "2016-2020\n Heritage"),
                         levels = c("2000-2004\n News21",
                                    "2016-2020\n Heritage")))

fraud %>%
  group_by(`VBM Status`, Period) %>%
  summarize(n = n(),
            count_mean = mean(1e6 * count/pop),
            count_sd   = sd(1e6 * count/pop)/sqrt(n)) %>%
  ggplot() +
  aes(x = Period,
      y = count_mean) +
  geom_linerange(aes(ymin = count_mean - count_sd,
                     ymax = count_mean + count_sd),
                 size = 2) +
  geom_linerange(aes(ymin = count_mean - 2 * count_sd,
                     ymax = count_mean + 2 * count_sd)) +
  labs(x = "",
       y = "number of fraud cases per million voters in most recent general election",
       color = "",
       title = "Figure 2.b. Fraud per voter declined sixty-three percent in vote
       by mail states and fifty-two percent in other states") +
  facet_wrap(~ relevel(factor(`VBM Status`), "Vote-by-mail state")) +
  coord_cartesian(ylim = c(0, 2), xlim = c(.25, 2.75), expand = FALSE)
```

```r
#Section 3
library("gsynth")
ballots_highest_office <-  read_csv("ballots_highest_office.csv")
voting_eligible_population <- read_csv("voting_eligible_population.csv")

election_project <-
  left_join(ballots_highest_office %>%
              gather("year", "num_ballots", -state),
            voting_eligible_population %>%
              gather("year", "num_voters", -state)) %>%
  mutate(year = as.numeric(year))

synthetic_control_plot <- function(treated_state, treated_year) {
  year_state_all <-
    expand.grid(YEAR = 2001:2020,
                STATE = setdiff(datasets::state.name,
                                setdiff(VBM, treated_state))) %>%
    mutate(`VBM Status` = ifelse(STATE == treated_state,
                                 "Vote-by-mail state",
                                 "Not a vote-by-mail state"))
  combined_data_all <-
    heritage %>%
    filter(YEAR > 2012) %>%
    left_join(eavs_2016, by = c("STATE" = "State")) %>%
    group_by(YEAR, STATE) %>%
    summarize(count = n()) %>%
    rbind(news21 %>%
            filter(YEAR <= 2012) %>%
            rename("STATE ABBREV" = "STATE") %>%
            left_join(tibble(STATE = datasets::state.name,
                             "STATE ABBREV" = datasets::state.abb)) %>%
            group_by(YEAR, STATE) %>%
            summarize(count = n())) %>%
    group_by(YEAR, STATE) %>%
    summarize(count = sum(count))

  fraud_year <-
    year_state_all %>%
    left_join(combined_data_all) %>%
    left_join(
      rbind(election_project %>% mutate(year = year + 1),
            election_project %>% mutate(year = year + 2)),
      by = c("YEAR" = "year", "STATE" = "state")) %>%
    mutate(count = ifelse(is.na(count), 0, count),
           `VBM` = relevel(factor(
             ifelse(`VBM Status` == "Vote-by-mail state",
                    "VBM", STATE)), "VBM")) %>%
    group_by(YEAR, `VBM`) %>%
    summarize(count = sum(count),
              num_ballots = sum(num_ballots),
              num_voters = sum(num_voters)) %>%
    mutate(Period = factor(ifelse(YEAR <= 2012, "Before (News21)",
                                  "After (Heritage Foundation)"),
```

```
                          levels = c("Before (News21)",
                                     "After (Heritage Foundation)")))

  df_synth_all <-
    fraud_year %>%
    ungroup() %>%
    transmute(Y = count,
              X1 = num_ballots,
              X2 = num_voters,
              state = as.character(`VBM`),
              year = (YEAR - 2000),
              state.num = as.numeric(`VBM`),
              T1 = ifelse(state == "VBM" & year > 11, 1, 0))

  gsynth.out <- gsynth(Y ~ T1 + X1 + X2, data = df_synth_all, parallel = FALSE,
                       estimator = "mc", index = c("state.num","year"), se = TRUE,
                       nboots = 500, r = c(0, 5), CV = TRUE, force = "two-way")

  gsynth_plot_data <-
    gsynth.out$att.boot %>%
    as_tibble() %>%
    mutate(year = 1:20) %>%
    gather("sample", "att", -year) %>%
    group_by(year) %>%
    summarize(n    = n(),
              mean_att = mean(att),
              mean_att_l95   = quantile(att, .025),
              mean_att_u95   = quantile(att, .975),
              mean_att_l50   = quantile(att, .25),
              mean_att_u50   = quantile(att, .75)) %>%
    left_join(df_synth_all %>% filter(state == "VBM"))

  gsynth_plot_data %>%
    ggplot() +
    aes(x = year) +
    geom_ribbon(aes(ymin = Y - mean_att_l95, ymax = Y - mean_att_u95),
                alpha = .25) +
    geom_ribbon(aes(ymin = Y - mean_att_l50, ymax =Y - mean_att_u50),
                alpha = .5) +
    geom_line(aes(year, Y, linetype = outcome),
              data = rbind(gsynth_plot_data %>%
                             transmute(year, Y = Y - mean_att,
                                       outcome = paste("synthetic", treated_state)),
    gsynth_plot_data %>% transmute(year, Y = Y, outcome = treated_state)) %>%
      mutate(outcome = relevel(factor(outcome), treated_state))) +
    theme(legend.position = "bottom") +
    geom_vline(xintercept = treated_year) +
    scale_x_continuous(breaks = c(5, 10, 15, 20),
                       labels = c(2005, 2010, 2015, 2020)) +
    labs(linetype = "", x = "", y = "number of fraud cases")
}

set.seed(100)
```

```r
washington <- synthetic_control_plot("Washington", 11)
colorado <- synthetic_control_plot("Colorado", 13)

washington +
  labs(title = "Figure 3.a. Washington would have eighty more fraud cases had it
  not instituted its vote-by-mail policy (± ten cases).") +
  annotate("text", x = 11.25,
           y = 40,
           label = "Washington begins\n vote by mail",
           hjust = 0,
           size = 8) +
  coord_cartesian(ylim = c(-.5, NA), expand = FALSE)

colorado +
  labs(title = "Figure 3.b. Colorado would have one less fraud case had it not
  instituted its vote-by-mail policy (± two cases).") +
  annotate("text", x = 13.25,
           y = 5,
           label = "Colorado begins\n vote by mail",
           hjust = 0,
           size = 8) +
  coord_cartesian(ylim = c(-.1, NA), expand = FALSE)
```