Online Appendix

Intended for online publication only.

Contents

S.1 Sensitivity of Potential Case Rate Calculation

Throughout the paper, we estimate the number of potential cases of fraud using the rate of positive cases as a share of cases we can confirm either way and multiplying this by the number of potential cases. This calculation assumes that the cases we cannot confirm have a similar number of positives as the cases we can confirm. We cannot directly confirm this assumption. Still, we can rule out that the cases we fail to confirm are clearly different from the cases we can confirm in ways that would make them much more likely to be positive if we could confirm them.

To assess the possibility that the unconfirmed cases are different in important, observable ways, we estimate the rate of positive cases after accounting for observable characteristics of the case that may relate to the likelihood that the case is positive. We use logistic regressions of a flag for positive cases on a set of covariates, and calculate the average predicted probability of a positive for each potential case including the cases we cannot confirm either way. In each regression, an observation is a potential link, meaning that voters can be linked to multiple death records and some are in this analysis.

Table S.1 reports our estimates. In column 1, we report the positivity rate using the simple approach we use in the paper. In our regression framework, this is equivalent to on constant-only regression—assuming that all cases have an equal probability of being positive regardless of their characteristics. In column 2, we report estimates after relaxing this assumption, instead calculating a probability that a case is positive for each match type. We categorize the matches into five categories: exact name match; first and last name match but middle initial is missing in both records; first and last name match but middle initial is missing in one record; first and last name match but middle initials are different; last name matches but first name is slightly different. This adjustment does not meaningfully change our estimate of the positivity rate.

Column 3 accounts for the population of the county in which the person lived and voted. Since we are more likely to find positive—probably false positive—cases in counties with many people, adjusting for the county population could change our expected positivity rate if the unconfirmed and confirmed cases came from different counties. In column four, we adjust for the commonness of the decedent's last name, suspecting that common last names also increase the rate of false positives. We find that both of these adjustments are not consequential.

In column 5, we adjust for the availability of Social Security Death Index (SSDI). People born prior to 1936 or who died after 2014 may not be listed in the SSDI. When we adjust for this, our positivity rate goes up modestly. While we cannot directly translate this estimate

Table S.1 – Sensitivity of Plausible Case Rate Calculation.

	Plausible Cases/Potential Cases				
	(1)	(2)	(3)	(4)	(5)
	0.0048	0.0047	0.0046	0.0047	0.0061
Controls					
Match Type Dummies	No	Yes	Yes	Yes	Yes
Log(Deaths in County)	No	No	Yes	Yes	Yes
Log(Last Name Freq in Death Records)	No	No	No	Yes	Yes
SSDI Records Availability Dummy	No	No	No	No	Yes

Each cell reports an estimate of the share of plausible links that would be potential links. Estimates are average predicted probabilities from logistic regressions. Each regression regresses a dummy variable for a potential case on covariates expected to predict potential cases. Regressions are estimated using cases where the scraper finds definitive evidence of a potential case or rules the case out. The share of plausible links is estimated by using the regression to extrapolate to the cases the automated searching algorithm cannot classify.

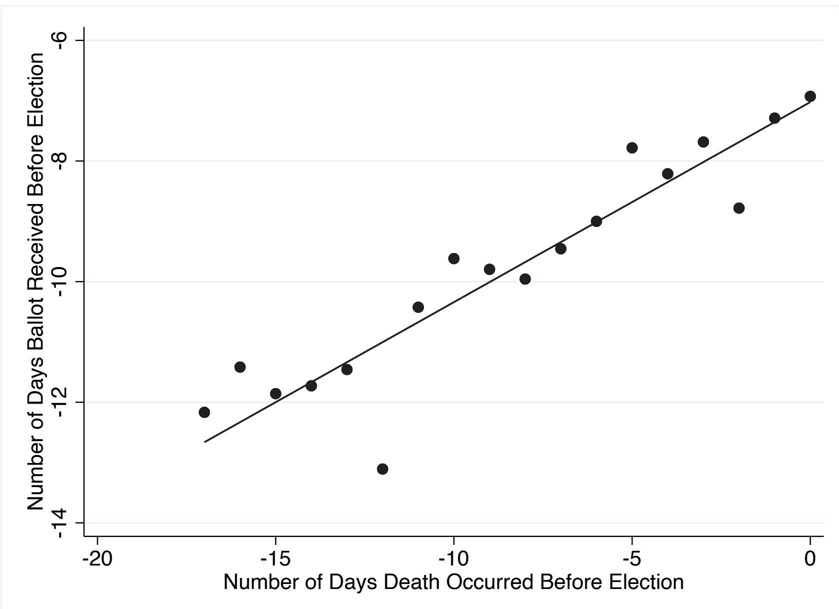
into a number of voters, we can approximate how these differences would change our main point estimate by inflating the rate we use for imputation by $\frac{0.0061}{0.0048}$ from columns 5 and 1. This would increase our point estimate from 53 to 68.

In total, Table S.1 tells us that our simple method of estimating the rate of plausible cases produces similar results as other methods that explicitly adjust for differences between the cases we can confirm and those we cannot.

S.2 Additional Results for Deaths Close to Election Day

If most ballots cast in the names of people who die shortly before an election are legitimate, then Washington should also receive the ballots of people who die earlier before those who die later. We check this using all deaths within 18 days of an election that match to a voter record.[33] We limit our analysis to death records that match a voter record on first name, last name, middle initial, county of residence, age, and gender. Figure S.1 confirms that ballots cast in the name of people who die two weeks before the election are received by the county much earlier on average than those cast by people who die closer to election day.

Figure S.1 – Ballots of People Who Die Earlier are Received Earlier. The horizontal axis plots the number days before the election that an individual dies. Each point represents the average of people who died a certain number of days before an election. The vertical axis plots the average number of days before the election that ballots are received from individuals who die on that day.



[33]We limit this analysis to the final 18 days before Election Day because that is the day that most voters begin to receive their ballots, and voting rates are very low before that.