Data Descriptor

Cast vote records: A database of ballots from the 2020 U.S. Election[*]

Shiro Kuriwaki[1,†,*], Mason Reece[2,†], Samuel Baltz[2], Aleksandra Conevska[3], Joseph R. Loffredo[2], Can E. Mutlu[3], Taran Samarth[1], Kevin E. Acevedo Jetter[2], Zachary Djanogly Garai[2], Kate Murray[2], Shigeo Hirano[4], Jeffrey B. Lewis[5], James M. Snyder, Jr.[3], and Charles H. Stewart, III[2]

[1]Yale University, Institution for Social and Policy Studies, New Haven, CT, 06511, USA
[2]Massachusetts Institute of Technology, Department of Political Science, Cambridge, MA, 02139, USA
[3]Harvard University, Department of Government, Cambridge, MA, 02138, USA
[4]Columbia University, Department of Political Science, New York, NY, 10027, USA
[5]University of California Los Angeles, Department of Political Science, Los Angeles, CA, 90095, USA
[*]Corresponding author: shiro.kuriwaki@yale.edu
[†]These authors contributed equally to this work.

ABSTRACT

Ballots are the core records of elections. Electronic records of actual ballots cast (*cast vote records*) are available to the public in some jurisdictions. However, they have been released in a variety of formats and have not been independently evaluated. Here we introduce a database of cast vote records from the 2020 U.S. general election. We downloaded publicly available unstandardized cast vote records, standardized them into a multi-state database, and extensively compared their totals to certified election results. Our release includes vote records for President, Governor, U.S. Senate and House, and state upper and lower chambers – covering 40.7 million voters in 20 states who voted for more than 2,121 candidates. This database serves as an uniquely granular administrative dataset for studying voting behavior and election administration. Using this data, we show that in battleground states, 1.9 percent of solid Republicans (as defined by their congressional and state legislative voting) in our database split their ticket for Joseph Biden, while 1.0 percent of solid Democrats split their ticket for Donald Trump.

Background & Summary

Ballots are the core records of elections. While *sums* of ballots for individual candidates are reported regularly at geographic levels (King et al. 1997; Ansolabehere, Palmer, and Lee 2014; Baltz et al. 2022; Voting and Election Science Team 2024), rarely are *individual* ballots made available.

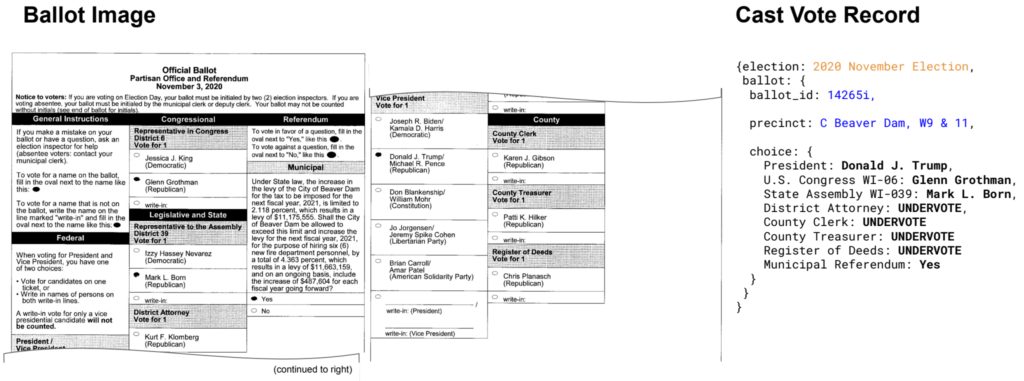
In paper-based elections, three sources of records at the individual level have been available to some researchers and litigators. The first is the actual, marked paper ballot. The second is the electronic scan of the paper ballot, often referred to as a *ballot image*. The third is an electronic record of the machine's interpretation of that scanned record, called *cast vote records* (CVRs).[1] CVRs are not the ultimate basis of an election, but among the three sources of data, it is the only source that is electronically transmittable (Figure 1). CVRs should directly reproduce vote totals produced by ballot tabulators.

In this study, we introduce a dataset of CVRs representing 40.7 million voters. Unlike certified election results — typically made available by state election officials — CVRs are rarely

[*] This document is a data descriptor for dataset https://doi.org/10.7910/DVN/PQQ3KV, V1 (July 2024).

[1] The National Institute of Standards and Technology (NIST) describes cast vote records as follows: "A CVR is an electronic record of a voter's selections, with usually one CVR created per sheet (page) of a ballot. Election results are produced by tabulating the collection of CVRs, and audits can be done by comparisons of the paper ballots or paper records of voter selections against the CVRs." (Wack 2019).

Figure 1. Cast Vote Record Example. An example of an actual ballot image (left) and the authors' representation of the associated cast vote record (right) in Wisconsin. Blank marks are recorded as an undervote.



Ballot Image

Official Ballot
Partisan Office and Referendum
November 3, 2020

Notice to voters: If you are voting on Election Day, your ballot must be initialed by two (2) election inspectors. If you are voting absentee, your ballot must be initialed by the municipal clerk or deputy clerk. Your ballot may not be counted without initials (see end of ballot for initials).

General Instructions

If you make a mistake on your ballot or have a question, ask an election inspector for help (absentee voters: contact your municipal clerk).

To vote for a name on the ballot, fill in the oval next to the name like this: ●

To vote for a name that is not on the ballot, write the name on the line marked "write-in" and fill in the oval next to the name like this: ●

Federal

When voting for President and Vice President, you have one of two choices:

- Vote for candidates on one ticket, or
- Write in names of persons on both write-in lines.

A write-in vote for only a vice presidential candidate **will not be counted.**

President / Vice President

Congressional

Representative in Congress District 6 Vote for 1

○ Jessica J. King (Democratic)

● Glenn Grothman (Republican)

○ write-in:

Legislative and State

Representative to the Assembly District 39 Vote for 1

○ Izzy Hassey Nevarez (Democratic)

● Mark L. Born (Republican)

○ write-in:

District Attorney Vote for 1

○ Kurt F. Klomberg (Republican)

Referendum

To vote in favor of a question, fill in the oval next to "Yes," like this ●.
To vote against a question, fill in the oval next to "No," like this ●.

Municipal

Under State law, the increase in the levy of the City of Beaver Dam for the tax to be imposed for the next fiscal year, 2021, is limited to 2.118 percent, which results in a levy of $11,175,555. Shall the City of Beaver Dam be allowed to exceed this limit and increase the levy for the next fiscal year, 2021, for the purpose of hiring six (6) new fire department personnel, by a total of 4.363 percent, which results in a levy of $11,663,159, and on an ongoing basis, include the increase of $487,604 for each fiscal year going forward?

● Yes

○ No

(continued to right)

Vice President Vote for 1

○ Joseph R. Biden/ Kamala D. Harris (Democratic)

● Donald J. Trump/ Michael R. Pence (Republican)

○ Don Blankenship/ William Mohr (Constitution)

○ Jo Jorgensen/ Jeremy Spike Cohen (Libertarian Party)

○ Brian Carroll/ Amar Patel (American Solidarity Party)

○ ______ / write-in: (President) ______ write-in: (Vice President)

○ write-in:

County

County Clerk Vote for 1

○ Karen J. Gibson (Republican)

○ write-in:

County Treasurer Vote for 1

○ Patti K. Hilker (Republican)

○ write-in:

Register of Deeds Vote for 1

○ Chris Planasch (Republican)

○ write-in:

Cast Vote Record

```
{election: 2020 November Election,
 ballot: {
  ballot_id: 14265i,

  precinct: C Beaver Dam, W9 & 11,

  choice: {
    President: Donald J. Trump,
    U.S. Congress WI-06: Glenn Grothman,
    State Assembly WI-039: Mark L. Born,
    District Attorney: UNDERVOTE,
    County Clerk: UNDERVOTE,
    County Treasurer: UNDERVOTE
    Register of Deeds: UNDERVOTE
    Municipal Referendum: Yes
  }
 }
}
```

centralized at the state level. Instead, CVRs are often produced as a byproduct of the tabulation process conducted at the sub-state level and retained by local election officials.

Following the November 3, 2020 U.S. general election, local election officials saw a surge in requests for cast vote records by anonymous constituents. Some states saw a four to five-fold increase in records requests between 2020 and 2022 (Green 2024; Leingang 2022). Owing to the work of election administrators in responding to these requests, and to O'Donnell (2023) for posting those unprocessed records online, researchers now have access to an unprecedented quantity of cast vote records.[2] We standardized the publicly available CVRs into a single database, and extensively compared them to official election results.

Our final dataset has two main audiences. The first audience are those in political science, economics, sociology, and other related fields who study electoral behavior. This data will allow researchers to study a wide variety of electoral phenomena. In particular, it will allow researchers to measure important aspects of voting behavior much more accurately than is possible using aggregate election returns, surveys or polls. The CVRs are at the individual level and record voters' actual choices across multiple offices. For example, the CVRs include choices for state legislative races. These are rarely included in surveys because researchers typically focus on top-of-the-ticket races.[3]

One important electoral phenomenon is split-ticket voting, that is, voting for candidates with different partisan affiliations across races. For example, aggregate-level election data can tell us how many votes Donald Trump received in a state and how many voted for each of the Republican candidates for offices further down the ballot, but it cannot tell us how many voters who voted for Donald Trump for President *also* voted for the Republican for all of the other offices (Burden and Kimball 2009; Malzhan and Hall 2024). Researchers can use CVRs to measure this type of behavior precisely, and at different levels of government, e.g., national level, state level, and across levels (Reece et al. 2024; Conevska et al. 2024; Kuriwaki 2023). Researchers can also use CVRs to count the number of voters who vote for Democratic candidates and also vote for the progressive position in referendums (Dubin and Gerber 1992; Gerber and Lewis 2004).

The data include geographic identifiers (counties and often precincts). Researchers can also infer geography from the CVRs themselves, since they reveal which voters were eligible to cast

[2] O'Donnell and his collaborators originally collected the CVRs in order to investigate the validity of the 2020 presidential election (Bloomberg Technology 2022). For a discussion of such claims, see Grimmer, Herron, and Tyler (2024)

[3] Moreover, estimates of vote choices in down-ballot elections can result in especially large measurement errors because respondents are less likely to correctly recall who they supported in down-ballot elections. Also, such estimates may have especially large sampling errors because only a few respondents in a typical nationally representative sample will have participated in any particular down-ballot contest such as those for state legislative seats.

a vote in which constituencies. One application would be to measure which state legislative or congressional districts have an especially high share of split-ticket voters. To the degree that these voters can be viewed as “swing” or “persuadable” voters, while straight-ticket voters can be viewed as “core” or “loyal” partisan voters, this measure could be used to test theories of resource allocation and campaign strategies.

Researchers can of course merge the CVR data with other information about the candidates and contests, such as incumbency status, patterns of campaign spending, and candidate attributes, such as race, ethnicity, gender, ideology, and/or experience (Dowling, Miller, and Morris 2024). With this extra data in hand, researchers can begin to study a wide variety of phenomena, including which voters split their tickets and in which ways as a function of the available choices. They can also study the degree to which split-ticket voting favors incumbents or the candidate who has a fundraising advantage. Researchers can also merge the CVR data with precinct-level demographic and socio-economic information to explore relationships between split-ticket voting and these types of variables. For example, does ticket splitting vary with the types and amounts of media (e.g., local newspapers) available in an area (DeLuca 2024)?

The CVR data could be used to study other electoral phenomena as well. Consider for example roll-off, where voters cast ballots but choose to abstain in particular races (also called undervoting). Using the CVRs, researchers can measure this behavior accurately. They can then investigate the types of races where this behavior is more prevalent, and the types of voters (e.g., straight-ticket Democrats, straight-ticket Republicans, or ticket-splitters) that are more likely to roll-off, as well as the interaction between contest and voter types. One final example is voting for candidates of minor parties, such as the Libertarian Party and the Green Party. Since the CVR database has millions of records, it contains many thousands of records of individuals who supported minor parties that received a small fraction of the vote (Herron and Lewis 2007). The typical survey is less useful for studying this type of behavior due to sample size.

The second audience is those in election law, forensics, and administration, who seek to study the integrity of the electoral process (Adler and Hall 2013; Bernhard et al. 2017). Cast vote records are of interest to scholars of attitudes towards election integrity because voters’ mistrust of the vote counting process, when it exists, often revolves around the ballots (Gerber et al. 2013a; Atkeson et al. 2023; Jaffe et al. 2023). Ballot-level data have been used in studies of election administration to help explain seemingly surprisingly election results (Wand et al. 2001; Bafumi et al. 2012). Our dataset can also contribute to debates around the tradeoff between transparency and privacy (Biggers et al. 2023; Gerber et al. 2013b; Williams, Baltz, and Stewart 2024; Kuriwaki, Lewis, and Morse 2023).

Finally, our data has implications that go beyond the particular states in this release, or even the 2020 election in particular. The November 2020 election was a turning point in U.S. politics, where the administration of elections became an overtly partisan issue. The conduct and administration of any other future elections now risks being politicized. Self-enforcing the legitimacy of elections requires election administrators, data scientists, and social scientists to work together upon a common understanding of election technology. We hope that our data release serves as a standard for future work in this area of growing interest.

Methods

We downloaded raw CVRs from `votedatabase.com`.[4] As described in O’Donnell (2023), this website provides access to the raw electronic file containing CVRs that were acquired by numerous citizens via open records requests.

We approach the raw data cautiously. CVRs may not include all of the ballots cast in an election for a number of reasons. Even within a single electoral jurisdiction, it is possible that some valid ballots are cast and tabulated in a way that creates CVRs while others are not. Sometimes ballots that are held aside for manual adjudication, such as provisional or damaged ballots, are

[4] Accessed on July 1, 2024.

not scanned through tabulators, and therefore a cast vote record is not created for such ballots. We can increase our confidence that CVR files posted are complete, uncorrupted, and genuine by comparing vote tallies produced using the downloaded CVR data to the official reports of vote totals from the same jurisdictions.[5]

We start with data that O'Donnell (2023) and his collaborators obtained and subsequently uploaded to `votedatabase.com`. O'Donnell (2023) reports that citizens requested CVRs according to their state's open record law guidelines, noting that all records were "obtained through these valid public records requests" (p.10). He reports that requests were sent to "nearly all counties in all states,"[6] and that 23 states responded that they did not have any records relevant to the request that could be provided under the state's open records law.[7] Indeed, the availability of CVRs is also limited by state law and executive order, with states including North Carolina, South Carolina, and New York shielding the CVR from open records requests.[8] In all, the CVRs in the data presented here originate from counties in 27 states and D.C. that were collected and made available by O'Donnell (2023).

Our standardization and validation proceeded in five steps. First, we downloaded the data files and standardized them so that their values were comparable across different voting machines and jurisdictions. While the website does provide some normalized versions of the raw data, we have chosen not to rely on these and instead independently process the raw data. We only considered files that were clearly CVRs in machine-readable form covering multiple offices and more than a handful of precincts. Standardization here entails that we identify the party affiliation of each candidate, code invalid votes consistently across jurisdictions, and standardize the formatting of candidate names. In the initial phase of the study, two groups conducted these pursuits independently, without awareness of each other's work. This gave us nearly independent measures of inter-coder reliability and reduced the possibility that a single coding error propagated to all counties.

About 15 percent of the counties had CVRs in non-rectangular formats such as JSON or XML. The remaining files were tabular files such as CSV or Microsoft Excel. Formats differed by the vendor of the machine (the main three being Dominion, Hart Intercivic, and ES&S), and the make of each machine.[9] We parsed these data with our own R and Python scripts, eventually normalizing all data into a long format where one row represents a single vote choice by a voter. This article releases the full codebase we developed.

Second, we assigned a cast vote record identifier to each voter, within a county or jurisdiction. Most times, this number indicates an individual voter – one anonymous voter gets one identifier for all their choices. In about 30 counties with ballots spanning multiple double-sided pages, each page was separated before it was scanned.[10] In about 20 of these counties, we used metadata to link pages into a single ID (see Appendix B). In the remaining 10, the records were irreversibly separated. This includes several counties in California (Los Angeles, San Francisco, San Bernardino, Ventura Counties). However, in the remaining counties, there is a one-to-one correspondence between the ID we assign and a single voter. Even in many counties with long ballots such as Maricopa, Arizona

[5] Our goal is *not* an audit of the election. CVRs are neither sufficient nor necessary to prove the election was valid. While CVRs are convenient representations of the paper ballot, they are simply not intended to be the official results of an election. A county could choose to hand-count some paper ballots and validly certify the election, without producing a cast vote record. On the other hand, even if the cast vote record were complete, ballots could have been tampered with beforehand. See also footnote 14.

[6] `votedatabase.com` did not limit their search to states battle ground states that Biden won. The database contains data from swing states like Wisconsin, Michigan, and Georgia, *as well as* solidly Democratic states, including California and New Jersey, and solidly Republican states, including Texas.

[7] O'Donnell (2023) reports that those states were Alabama, Connecticut, Hawaii, Indiana, Kansas, Louisiana, Maine, Massachusetts, Mississippi, Missouri, Montana, Nebraska, New Hampshire, New York, North Carolina, North Dakota, Oklahoma, South Carolina, South Dakota, Utah, Virginia, Washington, and Wyoming.

[8] See `https://commons.wikimedia.org/wiki/File:Ballot-foia.png` and Kuriwaki, Lewis, and Morse (2023).

[9] For estimates of the particular machine used in each county, see `https://verifiedvoting.org/verifier`.

[10] From the administrator's perspective, once the ballot itself is deemed valid, the identity of the voter is irrelevant in the counting process.

Name	Description
state	The name of the state that the ballot is from
county_name	The name of the county the ballot is from
cvr_id	A unique ID given to each ballot within a state-county
precinct_medsl	One field that indicates the precinct that can be matched to Baltz et al. (2022)
precinct_cvr	Concatenated precinct values as recorded in the raw data, for each precinct_medsl.
office	The name of the office the voter is choosing a candidate in
district	The district of the office the voter is choosing from
candidate	The name of the candidate the voter has selected in the office-district
party	A simplified name of the party of the candidate on the ballot
party_detailed	The full name of the party of the candidate on the ballot
magnitude	The number of candidates a voter could have chosen in this particular contest

Table 1. Dataset Variable Description

(with around 60 contests per ballot), the ballot record for each individual voter is preserved.

Third, we extensively checked the CVRs against other official sources of data, mainly the MIT Election Data and Science Lab's 2020 precinct-level returns (MIT Election Data and Science Lab 2022), as documented by Baltz et al. (2022). Baltz et al. (2022) was an ideal dataset to use as validation because it is at the precinct-level, it has standardized its formatting across states, it includes district-level as well as statewide contests, and features extensive documentation. We limited our attention to six offices: U.S. President, Governor, U.S. Senate, U.S. House, State Senate, and State House. The CVRs include votes for many other offices, including local administrative offices, school boards, and referendums. Other work analyzes these offices (Conevska et al. 2024; Reece et al. 2024) but we exclude them in our initial dataset release because we lack fully standardized official data to extensively validate against.

After attempting to find the best matching official result for all counties available on votedatabase.com, we only release counties with candidate-level discrepancies of 1 percent or less at the candidate-county level. This results in 352 counties in 20 states. The section "Technical Validation" discusses the discrepancy procedure in more depth.

Fifth, we extracted precinct information in the cast vote records and standardized them to match the MIT Election Data and Science Lab database of standardized precinct names. Precinct identifiers were often either names (e.g., "City of Madison Ward 1") or numeric codes (e.g., Precinct 35-001). These classifications are known to vary widely across jurisdictions and machines, with no national standard. We used fuzzy string matching and triangulation of vote counts to link the precincts, which we detail in Appendix A. There are 306 counties which we matched to the precinct level database.

Data Records

Our dataset includes 160,135,870 rows, each indicating a choice of a voter for a given contest. Our data is deposited in Dataverse at https://doi.org/10.7910/DVN/PQQ3KV (Kuriwaki, Reece, et al. 2024).

Variables Table 1 describes the variables in the data. Voters are uniquely identified by the state, county name, and CVR ID assigned by ourselves, with the exception of unmerged fragmented ballots described in the methods section. Contests are uniquely identified by the state, office, and district. In all offices but the State House in Arizona and West Virginia, the contests represented here are single-choice elections (a magnitude of one).[11] The names and values for our variables

[11] Arizona voters get 2 valid votes to elect their Representatives for the state's lower chamber, with 2 winners per district. Until 2020, some West Virginia voters had 2 or 3 valid votes for their lower chamber, in districts with 2 or 3 winners. West Virginia's lower chamber became a single member district system in 2022. It's state senate districts are represented

generally follow the naming convention in Baltz et al. (2022).

The `candidate` value is the name of the candidate that ballot was cast for, or it can be an undervote (`UNDERVOTE`), overvote (`OVERVOTE`), or write-in (`WRITEIN`). Undervotes refer to a blank choice for that contest or mark that the tabulator could not interpret, and overvotes refer to a voter marking more candidates than they could vote for in a given contest. Even though both types of votes are invalid, undervoting in particular is seen as a form of contest-specific abstention and is of interest to election scholars.

Third-party candidates and write-in candidates rarely win themselves, but the ideological orientation of voters who vote for them is of interest to researchers (Herron and Lewis 2007). When a candidate is listed on the ballot with a registered party, they are listed with their party affiliation in the `party_detailed` variable (such as Libertarian, Green, or "No Party Affiliation" in Florida). The `party_detailed` variable is set to missing for undervotes and overvotes. Jurisdictions vary in whether write-ins are reported separately or grouped together simply as write-in votes, and ballot access for third parties also varies by state. When the candidate is not listed on the ballot with a party, we record them as write-in candidates with no party affiliation (For more details, see Appendix C).

Our dataset also records the precinct of the cast vote record, as discussed in the Methods section. We provide two variables for precincts (Table 1). `precinct_medsl` is our matched precinct name, formatted to correspond exactly with those in Baltz et al. (2022). `precint_cvr` is a concatenation of the original precinct or precinct portion name as it appears in the cast vote record by the pipe character `|`. The concatenation occurs for every set of precincts for a given `precinct_medsl`. For example, if precinct `001` as defined in Baltz et al. (2022) contains two precinct portions `001A` and `001B`, and the cast vote records record the precinct portion, we give all voters in precinct `001` the value of `001A | 001B`.

We provide only the concatenated values to avoid facilitating the linkage of any particular voter to the vote that they cast. One concern with including precinct values in the cast vote record is that in rare instances, it may allow the unraveling of the secret ballot. In order to minimize the risk that the CVR data would allow for any voter revelation, we withhold the precinct information in counties where the conditions necessary for revelation to occur hold. To understand why election results can undo the secret ballot, suppose there is a precinct with only 3 voters in it. *Who* voted in U.S. elections is public via lists of voter rolls produced by election officials. Thus if all voters in the precinct voted for the same candidate and results were published at the precinct level, the vote choice for those 3 voters is revealed to anyone who has access to a voter list. Kuriwaki, Lewis, and Morse (2023) examines and outlines the extent of such *public revelation* and shows that CVRs reveal no more than granular election totals that are already published. Applying their method, we find that about 1 percent of our matched precincts, located in 32 counties of our released data, are ones in which such a revelation would be possible with the individual-level cast vote record but not with the precinct-level returns.[12] To be conservative, we remove all precinct values from all these 32 counties, i.e. not only in the revealed precinct, but all other precincts in the county.[13]

Geographic Coverage A total of 20 states are covered by our CVR data. Figure D.1 shows the geographic distribution of our data by state, and Table D.2 lists all counties in our released dataset. How does the coverage of our available data compare to the entire U.S.? Table 2 compares the characteristics of our collected samples with the entire state. A convenient cross-state metric to indicate the partisan tilt of our sample is the percentage of voters in our dataset that vote for the

by 2 senators, but they run on staggered terms.

[12] Even though precinct-level returns can also reveal vote choices, Kuriwaki, Lewis, and Morse (2023) notes that their revelations are ambiguous *if* they do not report undervotes. Because CVRs report undervotes, this can make some potential revelations more certain in a way that Baltz et al. (2022) cannot.

[13] Moreover, because the raw CVR files made available on `votedatabase.com` contain the unredacted precincts, our dataset does not make more revelations possible than they already were. If we had not redacted the individual precinct values, our standardization of the data would have made the revelations facilitated by the CVR data easier to undertake. We have done such a redaction.

Table 2. Comparison of Data Coverage to Entire State. *The first set of columns compares the two-party Biden voteshare in our data (CVR) and the entire state (Pop.). The second set of columns shows the total number of valid votes (excluding undervotes and overvotes)*

Democratic presidential candidate, Joseph R. Biden. Table 2 compares the percentage of Biden voters as a share of Biden and Trump voters in the dataset and the state(s) as a whole. In some states, like Ohio, the Biden lean of the CVR is within 1 percentage point of the statewide election result, even though we do not have CVRs of over two-thirds of the state's voters. Overall, 56 percent of our collection's Presidential voters are Biden voters (as a percentage of the two-party vote), while his two-party share nationwide was 52.3 percent (Federal Election Commission 2022).

Figure D.2 compares the set of counties based on population, race, and urban-ness. We take statistics from the 2020 decennial Census and compare our counties with the full set of counties by visualizing the density of the data across its range. Our counties tilt slightly towards populous counties (i.e., we are missing very small counties), as shown by the mismatch in the blue and clear densities in the figure. Table D.1 further indicates that our sample resembles the average county. The average county in our dataset does not differ from the average U.S. county in terms of the fraction racial minorities, age, and home-ownership by more than 2 percentage points. However, our average tends to be more urban.

Technical Validation

We extensively analyzed each county's files and compared them with official results, as described in the previous section.

Validation In general terms, we define a discrepancy as any difference between the summed total of CVRs cast for candidates in the CVRs we processed from a county and the certified results published by that county. More precisely, we compute the discrepancy in a county k by the percentage

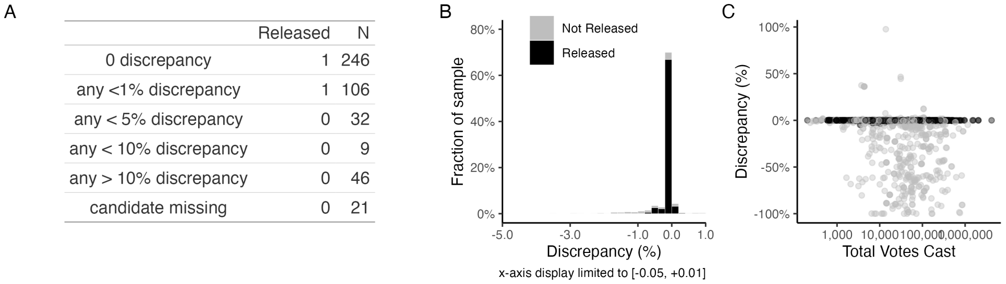
$$\text{discrepancy}_k = \max{}_{j \in \mathbb{I}_k} \left\{ \left| \frac{N_j^c - N_j^v}{N_j^v} \right| \right\}, \tag{1}$$

where N^c is the CVR vote count, N^v is the official vote count, j indexes candidates in a county, $|\cdot|$ is the absolute value function, and $\mathbb{I}_k$ is the set of all Republican, Democratic, and Libertarian candidates in the six offices in county k. We limited our validation to these three parties because affiliations for other choices can be reported in different ways by the county or voting machine, leading to mismatches even when the count is correct. We relied on Baltz et al. (2022) as our official vote count, and collected official statements of the vote from counties directly where necessary.

We then release counties where the discrepancy is 1 percent or less. Figure 2 shows the distribution of discrepancies. In panel A, we binned the maximum error rates at the county-level, showing that 246 counties had no discrepancy across all six offices, and 106 had a candidate with discrepancies within 1 percent. Panel B shows the candidate-level discrepancies, focusing on the range of -5 to 1 percent. We see that almost all the discrepancies are within 0.2 percent (and most are exactly zero). Some data points with 0 discrepancy are nevertheless not released and shown in light gray. This is because those counties have zero discrepancies in some offices but a discrepancy larger than 1 percent in other offices. Panel C shows the correlates of discrepancy by population. The discrepancies tend to be dispersed across large and small jurisdictions.

In addition to computing discrepancies at the county-level, we also conducted a precinct-level validation. Our matching procedure described previously produced 21,305 precincts in 306 counties that could be matched to standardized precinct names in Baltz et al. (2022). Figure 3

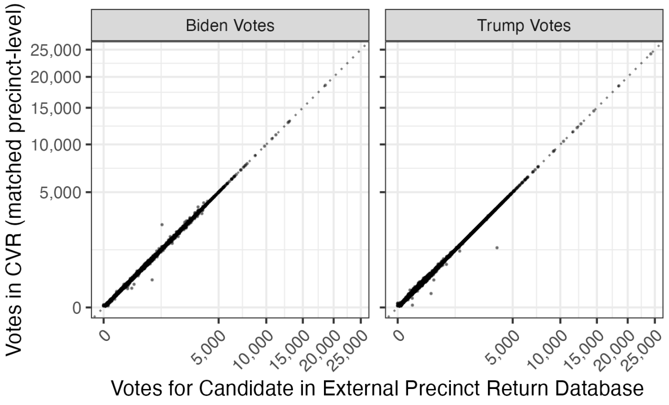
Figure 2. Error Rates. *(A) number of counties by discrepancy. (B) distribution of errors at the candidate level, limited to the neighborhood of 1 percent. (C) relationship between differences in total votes.*



A

	Released	N
0 discrepancy	1	246
any <1% discrepancy	1	106
any < 5% discrepancy	0	32
any < 10% discrepancy	0	9
any > 10% discrepancy	0	46
candidate missing	0	21

Figure 3. Precinct Level Validation. *Comparisons of vote totals for each Presidential candidate at the precinct level, with those from a precinct-level database (Baltz et al. 2022) on the x-axis and those from our cast vote record database (with approximately matched precinct) on the y-axis. Axes are shown on a square root scale.*



shows the alignment between the total votes for Presidential candidates in the CVR, per precinct, and the corresponding votes independently reported by the precinct-level dataset in Baltz et al. (2022). The match is not perfect due to the inclusion of counties with up to 1 percent discrepancies. Nevertheless, 19,129 precincts (out of 21,304 assigned) matched exactly, and 20,341 matched within 3 votes.

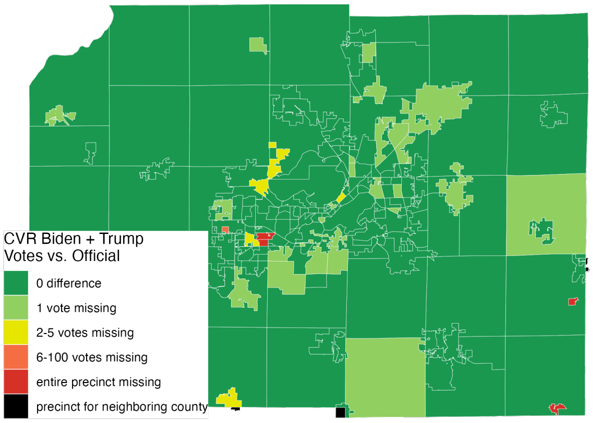
Reasons for Discrepancy Through our extensive validation, we found that reasons for divergence, in the rare cases where it happened, tended to fall into one of several categories.

First, in some counties, entire precincts were missing from the cast vote record data. This can happen if a county chooses to hand-count a batch of its ballots, or if the precinct's ballots are processed differently.

Second, in some counties the cast vote record data did not include votes cast by certain methods (vote by mail, in-person, provisional). For example, in Santa Rosa County, Florida and Cuyahoga County, Ohio, it appears that the mailed-in votes have not been included in the cast vote record export uploaded to O'Donnell (2023)'s website. We verified this by comparing our counts with the county's published vote totals broken out by vote method. This may occur if mailed and in-person votes are handled by a different tabulator.

Another type of known discrepancy is due to redactions done by counties to protect the integrity

Figure 4. Precinct-level Validation. An example from precinct-level matches in Dane County, Wisconsin



of the secret ballot, as we previewed in the Data Description section. Thirteen counties in Colorado and California provided CVRs where the vote choices of ballots in small ballot styles were removed. This practice is becoming increasingly common, and it is one way in which jurisdictions have tried to balance their dual roles of transparency and privacy. However, perfect redaction is known to be difficult because counties also need to report the total number of votes cast with perfect fidelity, and the redacted information can be backed out by triangulating the total election results. We do not attempt to unredact the CVRs.

There were many other counties with smaller discrepancies that could not be resolved with publicly available data. We believe one possible explanation for these minute discrepancies is the designation of provisional and disputed votes and ballot that were hand-tabulated. Our cast vote record may either exclude or include votes that were disputed but were later counted towards a major party candidate in the certified tally.[14] While the dispute resolution process is publicly recorded on video in many counties, we cannot link each resolved ballot to a record in our database.

The case of Dane County, Wisconsin (containing Madison) is illustrative in showing how several of these discrepancies can manifest in multiple ways. Figure 4 shows a precinct map of the county, colored by type of discrepancy. Two precincts were missing from the Dane CVR because they were held by the neighboring county. Wisconsin cities and villages, which conduct their own election tabulation, may be located near a county border and straddle two counties. In three other instances, we found that a cast vote record from a town actually belonged to the jurisdiction of a different county (and thus manually excluded them from the Dane CVR). In two other precincts in

[14] For example, Dodge County, Wisconsin's website cautions that

> "Please note Cast Vote Record (CVR) Reports are unofficial results from election night. These are the results the voting equipment tabulated on Election Day. The final, official canvass results posted on the Wisconsin Elections Commission's website for any state/federal races also include counted provisional ballots and other small adjustments. These adjustments are not tallied by, or in, the voting equipment, [but] rather through the County Board of Canvass process. The Cast Vote Record (CVR) Reports contain all data fields available in the ES&S Election Software. Also, please note that if a Municipal Clerk has accidentally corrupted their election data after printing their results tapes and electronically transferring the results into the County for a specific election, that data will not be able to be archived and therefore, would have no ballots to be read and included in the CVR Report."

(https://bit.ly/46eISNX, accessed July 1, 2024)

Dane County, scanning machine failures prompted a hand-count of the votes, with no ballot image or cast vote record present. In 25 other precincts, the cast vote record was only 1 vote short of the reported election result.[15] In this case, we contacted the Dane County clerk to resolve these issues, and obtained a letter written by the clerk's office that resolved our discrepancies (Dane County Clerk 2021). The issues fell into the one of the several categories discussed here.

We chose to limit our inquiry to county clerks to a minimum, given that many are already at or above capacity in their day-to-day duties (Ferrer, Thompson, and Orey 2024). We believe it is unlikely that the original collectors of the data tampered with these remaining cases before posting because discrepancies are small and inconsequential for the final outcome.

Usage Notes

In the remaining section, we illustrate how users can read in the data and analyze it. Although we use R as the running example, Python or any database-friendly programming language can read the dataset. We conclude with an example that studies the party loyalty of Republican and Democratic voters in their choice for President.

Reading in the Data We store our dataset in a parquet format. Parquet is a modern file storage format optimized for querying large datasets. It is partitioned by grouping variables, and it is columnar (so that users do not need to read in an entire row to extract a value from one column). Our dataset is prohibitively large to read and write in a plain-text format (20 Gb), but is compact and easy to read from in parquet (700 Mb). In R, we use the `arrow` package to query parquet files.[16]

The following command opens the dataset.

```r
library(tidyverse)
library(arrow)

ds <- open_dataset("cvrs")
```

Here, `"cvrs"` indicates the path to the top-level folder containing the parquet files downloaded from Dataverse. Our data is organized by county, nested within states. After unzipping the zip file in Dataverse[17], we see that `cvrs` has the following structure:

```
├── state=ARIZONA
│   ├── county_name=MARICOPA
│   │   └── part-0.parquet
│   ├── county_name=PIMA
│   │   └── part-0.parquet
│   ├── county_name=SANTA%20CRUZ
│   │   └── part-0.parquet
│   └── county_name=YUMA
│       └── part-0.parquet
...
── state=UTAH
│   └── county_name=SAN%20JUAN
```

[15] This should not be deemed as a case of election misconduct. The cast vote record, as we have made clear, is not meant to be the official data upon which electoral counts are computed. Instead, CVRs should be considered byproducts or an electronic trail of the ballot tabulator. See, for example, footnote 14

[16] For more information on how to read and write parquet files in R, see `https://r4ds.hadley.nz/arrow`. Parquet is also designed for usage in Python (`https://arrow.apache.org/docs/python/parquet.html`) and several other programming languages.

[17] Users can see the file hierarchy prior to download by using Dataverse's file preview feature.

```
│       └── part-0.parquet
└── state=WISCONSIN
    ├── county_name=BROWN
    │   └── part-0.parquet
    ├── county_name=KENOSHA
    │   └── part-0.parquet
    ├── county_name=PIERCE
    │   └── part-0.parquet
    └── county_name=WAUKESHA
        └── part-0.parquet
```

Because parquet is columnar, users will find it much faster to produce summary statistics of the data. Even though the code below counts some 161 million rows, it performs the count in one second on a personal laptop.

```
ds |> count(office) |> collect()
```

```
# A tibble: 6 x 2
  office              n
  <chr>           <int>
1 US PRESIDENT 40913498
2 US SENATE    18882854
3 US HOUSE     40731763
4 STATE SENATE 21531154
5 STATE HOUSE  38297179
6 GOVERNOR       562624
```

To perform this count, we used `count()` from `dplyr`, which totals the number of occurrences of each unique value in our `office` variable. We make use of R's pipe operator, `|>`, to pass our data objects forward onto subsequent operations we want to perform.

Finally, we must use the `collect()` command from `arrow` to extract the summary. All transformations before `count()` are *lazily-loaded*, meaning that they are not executed until needed. The arrow program combines the transformations internally in a way that avoids duplicative operations.

Extracting Summaries Users should use the combination of `state`, `office`, and `party` variable to identify candidates. The code below first limits to vote choices for President in Wisconsin ballots using `filter()` command, and counts the number of records for each candidate-party collection, sorted from most frequent to least.

```
ds |>
  filter(state == "WISCONSIN", office == "US PRESIDENT") |>
  count(candidate, party, sort = TRUE) |>
  collect()
```

```
# A tibble: 8 x 3
  candidate       party      n
  <chr>           <chr>  <int>
1 DONALD J TRUMP  REP   293283
2 JOSEPH R BIDEN  DEM   221356
3 JO JORGENSEN    LBT     6272
4 WRITEIN         <NA>    1412
5 UNDERVOTE       <NA>    1337
```

```
6 BRIAN T CARROLL OTH      874
7 DON BLANKENSHIP OTH      724
8 OVERVOTE        <NA>     493
```

For individual voters, use the cvr_id variable within a state and county. This ID is a numeric variable that is defined within counties. These numbers do not in any way indicate the time in which the ballot was cast, or the personal identity of the voter. The following code extracts the vote from the voter marked with the cvr_id of 23.

```
ds |>
  filter(state == "ARIZONA", county_name == "MARICOPA") |>
  filter(cvr_id == 23) |>
  select(county_name, cvr_id, office, district, candidate, party) |>
  collect()
```

```
# A tibble: 6 x 6
  county_name cvr_id office       district candidate               party
  <chr>        <int> <chr>        <chr>    <chr>                   <chr>
1 MARICOPA        23 US PRESIDENT FEDERAL  "JOSEPH R. BIDEN"       DEM
2 MARICOPA        23 US SENATE    ARIZONA  "KELLY, MARK"           DEM
3 MARICOPA        23 US HOUSE     003      "GRIJALVA, RAÚL"        DEM
4 MARICOPA        23 STATE SENATE 013      "KERR, SINE"            REP
5 MARICOPA        23 STATE HOUSE  013      "DUNN, TIMOTHY \"TIM\"" REP
6 MARICOPA        23 STATE HOUSE  013      "SANDOVAL, MARIANA"     DEM
```

This example shows that this voter split their ticket, voting for Democrats in the Presidential and Congressional race, while voting for one Republican candidate in state senate. However, further investigation into this voter's state senate district shows that it was uncontested.

```
ds |>
  filter(state == "ARIZONA", office == "STATE SENATE", district == "013") |>
  count(candidate, party_detailed) |>
  collect()
```

```
# A tibble: 8 x 3
  candidate       party_detailed     n
  <chr>           <chr>          <int>
1 "KERR, SINE"    REPUBLICAN     64192
2 "NOT QUALIFIED" <NA>            1852
3 "UNDERVOTE"     <NA>           34391
4 "BACKUS, BRENT" <NA>             145
5 "KERR SINE"     REPUBLICAN     29196
6 ""              <NA>             119
7 "WRITEIN"       WRITEIN          531
8 "OVERVOTE"      <NA>              17
```

We see that none of the ballots in State Senate district 13 were for a Democrat candidate, indicating that no Democrat ran in this district.

Application: Biden and Trump's Party Loyalty As our main exercise, we ask whether partisans — defined by their votes for Congress and state legislature — vote for their party's presidential candidate. Trump was a polarizing candidate. Election observers have wondered if Trump drew less support from Republican voters compare to Biden's support among Democratic voters. Some referred to these group of voters as "Never Trump Republicans.' '

For this analysis, we look at the counties in five battleground states which together decided the election: Wisconsin, Michigan, Georgia, Arizona, and Nevada.

```
ds_states <- ds |>
  filter(state %in% c("WISCONSIN", "MICHIGAN", "GEORGIA", "ARIZONA", "NEVADA"))
```

Recall that aggregate election results report how many votes Biden and Trump received, but unlike cast vote records, they do not reveal which of those votes came from Republicans and Democratic voters. Only cast vote records can classify voters into partisan types based on how they voted in all offices except President.

We first need to narrow down our data so that we only use voter-contest pairs in contests contested by a Democrat and a Republican. In other words, the voter needed to have a choice to vote for a Republican or Democrat.

```
ds_contested <- ds_states |>
  collect() |>
  # Contested contests
  filter(any(party == "REP") & any(party == "DEM"),
         .by = c(state, office, district)) |>
  # Ballots with Presidential vote
  filter(any(office == "US PRESIDENT"),
         .by = c(state, county_name, cvr_id))
```

The first `filter()` command in this code limits to vote choices for contested offices. For each state-office-district combination, we examine if there are any Republican candidates *and* any Democrats. Contests that do not meet this criteria are dropped. The second `filter()` command limits to ballots with a Presidential choice. This excludes fragmented ballots where the President and the rest of the ballot is separated. Both commands are done after `collect()` because the `arrow` package does not support group-specific filter commands as of version 16.1.0.

We now construct a dataset where each row is a single voter. We first create a dataset of Presidential votes:

```
## Voters based on President
ds_pres <- ds_contested |>
  filter(office == "US PRESIDENT") |>
  select(
    state, county_name,
    cvr_id, candidate,
    pres_party = party) |>
  mutate(pres = case_when(
    pres_party == "REP" ~ "Trump",
    pres_party == "DEM" ~ "Biden",
    pres_party == "LBT" ~ "Libertarian",
    candidate == "UNDERVOTE" ~ "Undervote",
    .default = "Other"))
```

Separately, we construct a dataset that classifies the same voters based on their non-Presidential vote choice. The variable `nonpres_party` is `Down-ballot Democrat` if the voter only votes for Democrats down-ballot (using the `all()` command) and it is `Down-ballot Republican` if the voter only votes for Republicans down-ballot.

```r
## subset to all-Dem voters based on everything except President
ds_D <- ds_contested |>
  filter(office != "US PRESIDENT") |>
  filter(all(party == "DEM"), .by = c(state, county_name, cvr_id)) |>
  distinct(state, county_name, cvr_id) |>
  mutate(nonpres_party = "Down-ballot Democrat")

## same subset, but for all-Rep voters
ds_R <- ds_contested |>
  filter(office != "US PRESIDENT") |>
  filter(all(party == "REP"), .by = c(state, county_name, cvr_id)) |>
  distinct(state, county_name, cvr_id) |>
  mutate(nonpres_party = "Down-ballot Republican")
```

Now we join voter's choices for President with their down-ballot choices. Because each row is now a single voter, we join one-to-one using `state`, `county_name`, and `cvr_id`. Voters who were not classified into Democrats or Republican, are, by construction, those who voted for some Democratic down-ballot candidates and Republican down-ballot candidates, or undervoted in some of these offices. We label them `Mixed`.

```r
ds_analysis <- ds_pres |>
  left_join(
    bind_rows(ds_D, ds_R),
    by = c("state", "county_name", "cvr_id"), relationship = "one-to-one") |>
  mutate(nonpres_party = replace_na(nonpres_party, "Mixed"))
```

Finally, we construct a cross-tabulation of this dataset using the base-R `xtabs()` function.

```r
xtabs(~ nonpres_party + pres, ds_analysis) |>
  addmargins()
```

```
                        pres
nonpres_party              Biden Libertarian   Other   Trump Undervote     Sum
  Down-ballot Democrat   2621937       11373    6878   26195      3443 2669826
  Down-ballot Republican   58887       23337    9212 3037114      8446 3136996
  Mixed                  1088384       69591   18509  698019     16581 1891084
  Sum                    3769208      104301   34599 3761328     28470 7697906
```

This table shows for example that among 2,669,826 solidly Democratic voters, 2,621,937 voted for Joe Biden. We can show cell counts in terms of proportions of the entire row, with the following operation:

```r
xtprop <- xtabs(~ nonpres_party + pres, ds_analysis) |>
  prop.table(margin = 1) |>
  round(3)

## add margins
N <- xtabs(~ nonpres_party, ds_analysis)

## reorder columns and append totals
xtprop[, c("Biden", "Trump", "Libertarian", "Undervote")] |>
  cbind(format(N, big.mark = ",")) |>
  kableExtra::kbl(format = "latex", booktabs = TRUE)
```

Table 3. **Party Loyalty in Five Battle Ground States**

	Biden	Trump	Libertarian	Undervote	
Down-ballot Democrat	0.982	0.010	0.004	0.001	2,669,826
Down-ballot Republican	0.019	0.968	0.007	0.003	3,136,996
Mixed	0.576	0.369	0.037	0.009	1,891,084

Table 3 table shows more clearly that the ticket splitting rate among solid partisans was on the order of 1 percent in this sample. Such small samples are almost impossible to detect in a survey. In contrast, 97 percent of solid Republicans stuck with their party's nominee, Trump, and 98 percent of solid Democrats stuck with Biden. Trump's party loyalty was only a percentage point smaller than Biden's.

A starker difference arises in the mixed group (those who vote for some Republicans and some Democrats down-ballot). Biden won this group of weak partisans by more than 20 points. Undervoting for President was low, less than 1 percent, in these group of voters. Down-ballot Republicans were only 0.2 percentage points more likely to undervote in the Presidential contest compared to down-ballot Democrats.

More can be done to examine if these results vary by state, county, or precinct. Future versions of this dataset can also include ballot measures and local candidates that give more context of these patterns.

Code availability

Code that we construct the dataset from downloaded files are available at https://github.com/kuriwaki/cvr_harvard-mit_scripts.

References

Adler, E. Scott and Thad E. Hall (2013). “Ballots, Transparency, and Democracy”. *Election Law Journal* 12.2, pp. 146–161. DOI: doi:10.1089/elj.2012.0179.

Ansolabehere, Stephen, Maxwell Palmer, and Amanda Lee (2014). *Precinct-Level Election Data, 2002-2012*. Draft version on Harvard Dataverse at https://doi.org/10.7910/DVN/YN4TLR.

Atkeson, Lonna Rae, Eli McKown-Dawson, M.V. Hood III, and Robert Stein (2023). “Voter Perceptions of Secrecy in the 2020 Election”. *Election Law Journal*, pp. 234–253. DOI: 10.1089/elj.2022.0064.

Bafumi, Joseph, Michael C. Herron, Seth J. Hill, and Jeffrey B. Lewis (2012). “Alvin Greene? Who? How Did He Win the United States Senate Nomination in South Carolina?” *Election Law Journal* 11.4, pp. 358–379. DOI: 10.1089/elj.2011.0137.

Baltz, Samuel, Alexander Agadjanian, Declan Chin, John Curiel, Kevin DeLuca, James Dunham, Jennifer Miranda, Connor Halloran Phillips, Annabel Uhlman, Cameron Wimpy, Marcos Zárate, and Charles Stewart III (2022). “American election results at the precinct level”. *Nature Scientific Data*. DOI: 10.1038/s41597-022-01745-0.

Bernhard, Matthew, Josh Benaloh, J Alex Halderman, Ronald L Rivest, Peter YA Ryan, Philip B Stark, Vanessa Teague, Poorvi L Vora, and Dan S Wallach (2017). “Public evidence from secret ballots”. *Electronic Voting: Second International Joint Conference, E-Vote-ID 2017, Bregenz, Austria, October 24-27, 2017, Proceedings 2*. Springer, pp. 84–109.

Biggers, Daniel R, Elizabeth Mitchell Elder, Seth J Hill, Thad Kousser, Gabriel S Lenz, and Mackenzie Lockhart (2023). “Can Addressing Integrity Concerns about Mail Balloting Increase Turnout? Results from a Large-Scale Field Experiment in the 2020 Presidential Election”. *Journal of Experimental Political Science* 10.3, pp. 413–425.

Bloomberg Technology (2022). *‘Raccoon Army’ Swamps Election Officials in Dubious Campaign to Disprove Results*. October 25, 2022.

Burden, Barry C and David C Kimball (2009). *Why Americans split their tickets: Campaigns, competition, and divided government*. University of Michigan Press.

Conevska, Aleksandra, Shigeo Hirano, Shiro Kuriwaki, Jeffrey B. Lewis, Can Mutlu, and James M. Snyder (2024). “Is All U.S. Politics National? Evidence from 2020 Cast Vote Records”. *Presented at Midwest Political Science Association Conference*.

Dane County Clerk (2021). *RE: November 2020 General Election Ballot Images*. Letter to Brian McGrath and Kyle Koenen on August 13, 2021.

DeLuca, Kevin (2024). “Editor’s Choice: Measuring Candidate Quality using Local Newspaper Endorsements”. *APSA Preprints*. DOI: 10.33774/apsa-2023-3qdmj-v3.

Dowling, Conor M., Michael G. Miller, and Kevin Morris (2024). “Can Voters Locate Copartisan Candidates in Nonpartisan Elections? Evidence from Cast Vote Records”.

Dubin, Jeffrey A and Elisabeth R Gerber (1992). *Patterns of voting on ballot propositions: A mixture model of voter types*. Tech. rep. California Institute of Technology, Division of the Humanities and Social …

Federal Election Commission (2022). *FEDERAL ELECTIONS 2020: Election Results for the U.S. President, the U.S. Senate and the U.S. House of Representatives*. Tech. rep. URL: https://www.fec.gov/resources/cms-content/documents/federalelections2020.pdf.

Ferrer, Joshua, Daniel M. Thompson, and Rachel Orey (2024). “Election Official Turnover Rates from 2000-2024”. *Bipartisan Policy Center*. URL: https://bipartisanpolicy.org/report/election-official-turnover-rates-from-2000-2024/.

Gerber, Alan S., Gregory A. Huber, David Doherty, and Conor M. Dowling (2013a). "Is there a secret ballot? Ballot secrecy perceptions and their implications for voting behaviour". *British Journal of Political Science* 43.1, pp. 77–102.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, and Seth J. Hill (2013b). "Do perceptions of ballot secrecy influence turnout? Results from a field experiment". *American Journal of Political Science* 57.3, pp. 537–551.

Gerber, Elisabeth R. and Jeffrey B. Lewis (2004). "Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation". *Journal of Political Economy* 112.6, 1364. DOI: 10.1086/424737.

Green, Rebecca (2024). "FOIA-Flooded Elections". *Forthcoming, Ohio State Law Journal*.

Grimmer, Justin, Michael C Herron, and Matthew Tyler (2024). "Evaluating a New Generation of Expansive Claims about Vote Manipulation". *Election Law Journal: Rules, Politics, and Policy*.

Herron, Michael C. and Jeffrey B. Lewis (2007). "Did Ralph Nader Spoil a Gore Presidency? A Ballot-level Study of Green and Reform Party Voters in the 2000 Presidential Election". *Quarterly Journal of Political Science* 2.3, pp. 205–226. DOI: 10.1561/100.00005039.

Jaffe, Jacob, Joseph Loffredo, Samuel Baltz, Alejandro Flores, and Charles Stewart III (2023). *What Effect do Audits Have on Voter Confidence?* Tech. rep. Working Paper.

King, Gary, Bradley Palmquist, Greg Adams, Micah Altman, Kenneth Benoit, Claudine Gay, Jeffrey B. Lewis, Russ Mayer, and Eric Reinhardt (1997). *The Record of American Democracy, 1984-1990*. Documentation at https://road.hmdc.harvard.edu/pages/road-documentation.

Kuriwaki, Shiro (2023). "Ticket Splitting in a Nationalized Era". *Working Paper*. URL: https://osf.io/preprints/socarxiv/bvgz3/.

Kuriwaki, Shiro, Jeffrey B. Lewis, and Michael Morse (2023). "The Still Secret Ballot: The Limited Privacy Cost of Transparent Election Results". *Working paper*. URL: https://arxiv.org/abs/2308.04100.

Kuriwaki, Shiro, Mason Reece, et al. (2024). *Cast vote records: A database of ballots from the 2020 U.S. Election, Harvard Dataverse*. URL: https://doi.org/10.7910/DVN/PQQ3KV.

Leingang, Rachel (Sept. 7, 2022). "Election activists are seeking the "cast vote record" from 2020. Here's what it is and why they want it." *Votebeat, September 7, 2022*. URL: https://arizona.votebeat.org/2022/9/7/23341640/cast-vote-record-data-ballot-tabulator-images.

Malzhan, Janet and Andrew B. Hall (2024). "Election-Denying Republican Candidates Underperformed in the 2022 Midterms". *Working Paper, Forthcoming, American Political Science Review*.

MIT Election Data and Science Lab (2022). *Precinct-Level Returns 2020 by Individual State*. Harvard Dataverse. DOI: https://doi.org/10.7910/DVN/NT66Z3.

O'Donnell, Jeffrey (2023). "The Fingerprints of Fraud: Evidence and Analysis of Multi-State Conspiracy to Defraud the 2020 General Election, Vol. 1". *Unpublished Manuscript Available Online*.

Reece, Mason, Gabrielle Péloquin-Skulski, Kate Murray, Joseph Loffredo, Kevin E Acevedo Jetter, Fernanda Gonzalez, Zachary Djanogly Garai, Alejandro Flores, Luka Bulic Braculj, Samuel Baltz, and Charles Stewart III (2024). "Hidden Partisanship in American Elections".

Voting and Election Science Team (2024). *Precinct-Level Election Results*. Last accessed 2024. URL: https://dataverse.harvard.edu/dataverse/electionscience.

Wack, John P (2019). "Cast Vote Records Common Data Format Specification Version 1.0". *National Institute of Standards and Technology*. DOI: 10.6028/NIST.SP.1500-103.

Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Michael C. Herron, and Henry E. Brady (2001). "The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida". *American Political Science Review* 95.4, pp. 793–810.

Williams, Jack R, Samuel Baltz, and Charles Stewart (2024). “Votes Can Be Confidently Bought in Some Ranked Ballot Elections, and What to Do about It”. *Political Analysis*, pp. 1–13.

Acknowledgements

We thank Luka Bulić Bračulj for additional work analyzing the data. We also thank Scott McDonnell of Dane County for answering our questions about Dane County, as well as other local election officials that answered our questions. We also thank Larry Moore for inspiration on the usage example. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2139841 (to T.S.).

Author contributions statement

S.K. drafted the manuscript; M.R., J.S., and J.B.L. were the principal maintainers of the database construction; A.C., C.S., J.B.L., J.R.L., J.S., M.R., S.H., S.K. contributed to the manuscript; A.C., C.M., J.B.L., J.R.L., J.S., K.A., K.M., S.H., S.K., M.R., T.S., Z.G., performed data processing, validation, investigation; J.B.L., J.R.L., J.S., M.R., S.K., S.B., T.S., wrote software; C.S., J.S., M.R., S.B., S.K. supervised the project. All authors reviewed the paper.

Competing interests

Authors declare no conflicting interests that might be perceived to influence the results and/or discussion reported in this paper.

Appendix

A Precinct Name Standarization

Linking CVR data to other precinct-level data sources is not straightforward. The released CVR data typically contains information about the voting precinct in which each ballot was cast. However, there is no agreed-upon standard for the labeling of precincts. The precinct labels used in the CVR data often do not match those used in the official precinct-level results published by the county or state. Further, CVRs sometimes provide more detailed geographic information (commonly called "subprecinct" information) that is not included in the official precinct-level tallies.

In order to make the CVR data more useful to researchers, we constructed a crosswalk from the precinct labels used in the CVR data to the precinct labels used in a widely-used freely-available national database of certified precinct-level 2020 General Election results published by the MIT Election Data and Science Lab (MEDSL) (see Baltz et al. 2022). Using that crosswalk, we have appended the MEDSL precinct label to our dataset.

We created this crosswalk by the following steps:

1. First, we aggregated the CVR data to the (sub)precinct level based on the precinct labels that they include.[18] We then attempted to match our CVR-based precinct records to the MEDSL precinct data. Working county-by-county, we matched CVR precincts to MEDSL precincts based upon the reported number of votes cast for Democratic and Republican candidates for US President, US Senate, US House, State Senate, and State House. For many counties, unique exact matches for every CVR precinct could be found among the MEDSL precincts. For those counties, this set of unique exact matches was used as the crosswalk between CVR and MEDSL precincts labels.

2. For counties in which unique exact candidate vote matches could not be found for every precinct, we transformed the CVR precinct names to more closely match the format of the MEDSL precinct names. We then disambiguated cases in which a given CVR precinct's candidate vote total exactly matched more than one MEDSL precinct using the edit distance between the transformed CVR precinct name and the MEDSL precinct name by selecting the potential match having the smallest edit distance. In cases in which no exact match based on the candidate vote totals existed, we established a linkage if the transformed CVR precinct name exactly matched the MEDSL precinct name.

3. For those counties containing precincts for which neither the candidate vote totals nor the transformed precinct names could be exactly matched across the CVR and MEDSL data, potential matches were identified by the smallest sum of absolute deviations in candidate votes and confirmed by manual comparison of the CVR and the MEDSL precinct labels. Where this was not possible, no linkage was established.

4. For some counties, we determined that the CVR precinct labels included subprecinct information meaning that the CVR-based precinct-level data was at a lower level of aggregation than that reported in the MEDSL data. In those counties, we transformed to CVR precinct names to remove the subprecinct information and re-aggregated to produce CVR precinct records at the same level of aggregation as the MEDSL precincts before constructing the crosswalk using the methods described above.

[18] For counties such as Los Angeles, California for which the CVRs do not provide precinct identifiers, we have not attempted to place the CVRs in MEDSL precincts. Such a linkage is not in general possible, though it might in be accomplished some cases by using the set of contests included on each ballot (sometimes called the "ballot style") to identify the precinct.

MEDSL precinct labels were not applied to records for which a precinct identifier was not present in the CVR data nor to CVR precincts for which neither the names nor the vote totals established a convincing linkage between the two data sources.

In four Michigan Counties (Alcona, Clinton, Gladwin, and Missaukee) few or none of the CVR precincts could be matched to MEDSL precincts using the method described. In these four counties, the inability to link CVR precincts to MEDSL precincts appeared to be the result of irregularities in the MEDSL data. Those issues may be resolved in the future.

Fifteen other counties contained between one and five CVR precincts that could not be linked to MEDSL precincts. All of the precincts in the remaining 296 counties for which CVR precinct information was available could be linked to MEDSL precincts. In total, 28,540 CVR unique (sub)precincts were linked to 23,467 MEDSL precincts.

B Fixing Fragmented Ballots

The counties in Table B.1 did not have a clear cast vote record identifier. We nevertheless paired records as one voter (one CVR ID) using a matching algorithm described in Algorithm 1. The algorithm relies on the user to supply two arguments, (1) a grouping column, `args.groupcol`, that is a condition for pages to belong to the same voter and (2) a target column, `args.targetcol`, in the CVR that can be used to match pages together. For example, if the column is the President, if two rows actually belong to the same voter, the pair must have an identical value for the `args.groupcol`, and one row will have a value for President and the other row will be blank for President.

Algorithm 1 Process Missing Values in Dataset

```
Initialize an empty set used_rows to keep track of processed rows.
Determine the total number of rows in the dataset, num_rows.
for each row current_row at index $i$ in the dataset do
    if $i$ is in used_rows then
        continue
    end if
    if current_row[args.targetcol] is missing a value then
        for each index $j$ in expanding spiral around $i$ do
            if $j \geq$ num_rows then
                break
            end if
            if $j$ is not in used_rows and current_row[args.groupcol] equals data[j, args.groupcol] and data[j, args.targetcol] is not missing then
                Merge current_row with the complementary row at index $j$ to form merged_row.
                Update data[i] with merged_row.
                Add $j$ to used_rows.
                break
            end if
        end for
    end if
end for
```

C Third Party Candidates and Write-ins

Designations for third parties vary widely by state. This is compounded by the different ways counties print party on the ballot, and the types of voting machines store and record information about third party candidates and write-ins.

Table B.1. Counties Where Algorithm 1 was used to re-connect pages

County	Group Column
Alameda, California	Ballot Type
Contra Costa, California	Ballot Type
Kings, California	Ballot Type
Merced, California	Ballot Style
Riverside, California	Ballot Type
San Benito, California	Ballot Type
San Mateo, California	Ballot Type
Sonoma, California	Ballot Type
Yuba, California	Ballot Type
Denver, Colorado	Ballot Type
Eagle, Colorado	Ballot Type
Routt, Colorado	Ballot Type
Gwinnett, Georgia	Ballot Type
Baltimore, Maryland	Ballot Style
Baltimore City, Maryland	Ballot Style
Montgomery, Maryland	Ballot Style
Prince George's Maryland	Ballot Style
Butler, Ohio	Ballot Type
Champaign, Ohio	BallotStyleID
Cuyahoga, Ohio	Ballot Style
Greene, Ohio	Ballot Type
Rhode Island	Ballot Style

Howie Hawkins was the Green Party's nominee for U.S. President but the Green Party did not get ballot access in some states. Among the states that we examined, the Green Party appears to not have had ballot access in Arizona, Georgia, Pennsylvania, and Wisconsin. Some jurisdictions, e.g., in Wisconsin, do not report out Hawkins as a specific candidate in either the cast vote record or the election returns; he is lumped into simply `WRITEIN`. In Georgia, the cast vote record and the election returns report Hawkins as its own candidate, but they are still write-ins.[19] In all these cases, the `party` value for Hawkins is a Write-in, not the Green Party. In contrast, New Jersey's CVRs records each and every write-in choice as a valid candidate.

It is worth noting that some of the data on `votedatabase.com` lost information about write-ins due to information loss from Excel to CSV. In some machines, notably the ES&S DS200 scanner, write-ins are scanned and can produce a cast vote record in Microsoft Excel. In these sheets, the write-in candidates are not transcribed into text but stored as an image file. However, O'Donnell instructed his collaborators to upload these data as a plain-text CSV file, which loses information in the image. In these cases, write-in votes are tracked as blank cells and excluded from our dataset. Our data, therefore, systematically misses write-in votes in many counties that use this format of cast vote record.

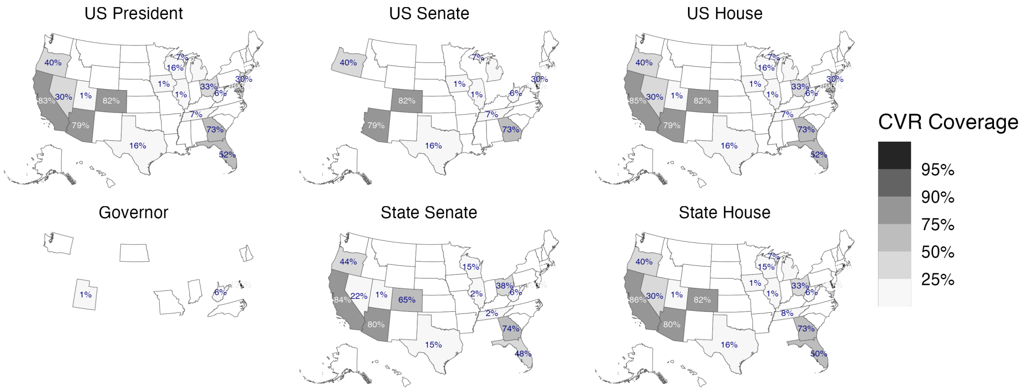
D Sample Characteristics

See Table D.2 for the list of counties included in our data release. In Delaware and Rhode Island, we count the whole state as one county given the format of the data.

Figure D.1 shows the states in our dataset and the coverage of its population for each office. The percentage values assigned for each state indicate the total number of ballots relative to the

[19] See `https://ballotpedia.org/Georgia_official_sample_ballots,_2020` for a sample ballot, and `https://bit.ly/3RYhL3A` for an example of how Hawkins is reported in an election return.

Figure D.1. Coverage of CVR data. *The percentage shows the fraction of total voters contained in the CVR for that state and office. States are not shown on the map if the office was not on the ballot. States have no percentage if CVRs were not available for that state.*



total count of votes reported in the entire state. For example, in Texas, we have around 15 percent of the votes for the President, Congress, and state legislature. There was no election for Governor in the state.

Figure D.2. Characteristics of County Data. *Comparison of the counties in our sample with all counties in the United States. Data all come from the 2020 Decennial Census. Plots show density plots for each group. The total population variable is shown on a square-root scale*.

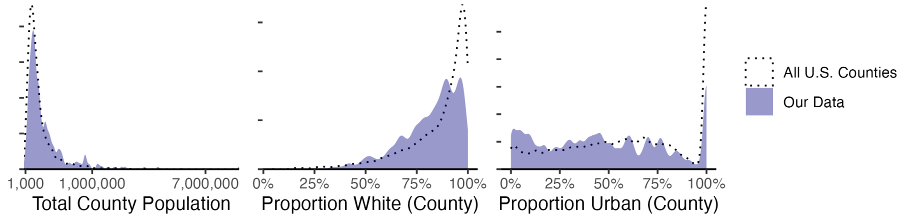


Table D.1. **Detailed Characteristics of County Data**. *Comparison of the counties in our sample with all counties in the United States. Data come from the 2020 Decennial Census. Column-pairs report key summary statistics across counties.*

	Mean		Median		Standard Dev.	
	Nation	CVR	Nation	CVR	Nation	CVR
Percent Urban	37.090	49.571	34.775	53.346	34.163	33.692
Percent White	75.312	71.735	82.148	75.051	19.939	16.848
Percent Black	8.678	9.873	2.238	3.756	13.954	13.387
Percent Hispanic	11.946	14.629	4.770	8.897	19.243	14.907
Percent Under 18	21.905	21.785	21.904	22.113	3.350	3.408
Percent Over 65	20.192	19.403	19.879	18.580	4.673	5.388
Percent Homeowners	28.529	27.054	29.033	27.030	4.425	4.532

Arizona
4 counties, 2,706,032 voters
Maricopa
Pima
Santa Cruz
Yuma

California
34 counties, 14,614,190 voters
Alameda
Amador
Contra Costa
Del Norte
El Dorado
Fresno
Glenn
Humboldt
Kern
Kings
Lake
Los Angeles
Marin
Mendocino
Merced
Nevada
Orange
Placer
Riverside
San Benito
San Bernardino
San Diego
San Francisco
San Luis Obispo
San Mateo
Santa Barbara
Santa Cruz
Shasta
Sonoma
Sutter
Tehama
Tuolumne
Ventura
Yuba

Colorado
57 counties, 2,698,653 voters
Adams
Alamosa
Archuleta
Bent
Broomfield
Chaffee
Cheyenne
Clear Creek
Conejos
Costilla
Crowley
Custer
Delta
Denver
Dolores
Douglas
Eagle
El Paso
Elbert
Fremont
Garfield
Gilpin
Grand
Gunnison
Hinsdale
Huerfano
Jefferson
Kiowa
Kit Carson
La Plata
Lake
Larimer
Lincoln
Logan
Mesa
Mineral
Moffat
Montezuma
Montrose
Morgan
Otero
Ouray
Park
Phillips
Pitkin
Prowers
Pueblo
Rio Blanco
Rio Grande
Routt
Saguache
San Miguel
Sedgwick
Teller
Washington
Weld
Yuma

Delaware
507,773 voters
(Statewide)

Florida
26 counties, 5,738,195 voters
Bradford
Broward
Calhoun
Citrus
Collier
Duval
Escambia
Flagler
Gulf
Hamilton
Hillsborough
Holmes
Lafayette
Lake
Manatee
Marion
Martin
Nassau
Okaloosa
Orange
Palm Beach
Pasco
St Johns
Sumter
Wakulla
Walton

Georgia
86 counties, 3,643,517 voters
Bacon
Barrow
Bartow
Ben Hill
Berrien
Bibb
Bleckley
Brantley
Bryan
Burke
Camden
Carroll
Catoosa
Charlton
Chatham
Chattooga
Cherokee
Clarke
Cobb
Colquitt
Columbia
Cook
Crisp
Dade
Dawson
Dekalb
Douglas
Early
Echols
Elbert
Emanuel
Fayette
Forsyth
Franklin
Gilmer
Glascock
Glynn
Gordon
Grady
Greene
Gwinnett
Hall
Haralson
Harris
Hart
Heard
Henry
Houston
Irwin
Jackson
Jasper
Lanier
Laurens
Lee
Lowndes
Lumpkin
Madison
Mcduffie
Mitchell
Morgan
Murray
Muscogee
Newton
Oconee
Paulding
Pierce
Pike
Polk
Pulaski
Putnam
Rabun
Richmond
Rockdale
Schley
Spalding
Talbot
Tattnall
Terrell
Thomas
Tift
Upson
Walker
Walton
Warren
Whitfield
Wilcox

Illinois
8 counties, 85,625 voters
Brown
Clinton
Hamilton
Jo Daviess
Monroe
Pike
Union
Wayne

Iowa
1 county, 11,244 voters
Dickinson

Maryland
21 counties, 2,420,605 voters
Allegany
Anne Arundel
Baltimore
Baltimore City
Calvert
Caroline
Carroll
Cecil
Charles
Dorchester
Frederick
Garrett
Harford
Howard
Kent
Prince George’s
Queen Anne’s
Somerset
Talbot
Washington
Wicomico

Michigan
10 counties, 403,539 voters
Alcona
Alger
Charlevoix
Clinton
Eaton
Gladwin
Iron
Mason
Missaukee
Washtenaw

Nevada
11 counties, 419,067 voters
Carson City
Churchill
Douglas
Elko
Lander
Lincoln
Lyon
Nye
Pershing
Washoe
White Pine

New Jersey
6 counties, 1,384,872 voters
Bergen
Cumberland
Gloucester
Monmouth
Passaic
Salem

Ohio
33 counties, 1,940,539 voters
Allen
Ashtabula
Auglaize
Belmont
Butler
Champaign
Clark
Clinton
Darke
Erie
Fayette
Franklin
Gallia
Greene
Hancock
Harrison
Highland
Hocking
Logan
Miami
Pickaway
Preble
Putnam
Richland
Ross
Seneca
Shelby
Trumbull
Tuscarawas
Van Wert
Wayne
Wood
Wyandot

Oregon
14 counties, 958,676 voters
Columbia
Coos
Douglas
Harney
Josephine
Klamath
Lincoln
Linn
Marion
Polk
Union
Wasco
Washington
Yamhill

Rhode Island
518,200 voters
(Statewide)

Tennessee
4 counties, 216,215 voters
Loudon
Pickett
Sevier
Williamson

Texas
28 counties, 1,800,681 voters
Andrews
Bosque
Collin
Comal
Cooke
Cottle
Crane
Ellis
Erath
Fort Bend
Grayson
Guadalupe
Hidalgo
Kendall
Lee
Montague
Navarro
Oldham
Orange
Parmer
Potter
Scurry
Smith
Stephens
Taylor
Walker
Washington
Wharton

Utah
1 county, 6,938 voters
San Juan

West Virginia
2 counties, 49,754 voters
Nicholas
Wood

Wisconsin
4 counties, 525,751 voters
Brown
Kenosha
Pierce
Waukesha

Table D.2. List of Counties Included