Supplemental Information

"Political Appointments and Outcomes in Federal District Courts"

Ryan Hübert
UC Davis

Ryan Copus
UMKC School of Law

For Online Publication

Replication code and data available at xxx.

A Data

The analysis in the paper utilizes two datasets: a dataset of district court cases from seven district courts and a dataset of circuit court cases from the Ninth Circuit. In this section, we provide additional details about these datasets.

A.1 Data Collection

We collected docket sheets for every case filed in the seven district courts in Washington, Oregon and California between 1995 and 2016. (Note: due to limits on our data collection process, we only have cases filed between late 1996 and early 2015 for the District of Oregon.) We also collected all docket sheets from the Ninth Circuit from 1996 to 2012. One of the authors was granted fee waivers from each court to access these records directly from the federal judiciary's electronic records system, PACER (see https://www.pacer.gov/).

A.2 Constructing the Dataset

To construct our dataset, we took the following steps.

Step 1. FJC IDB The core of our dataset consists of case records available in the FJC's Integrated Database (IDB, available at https://www.fjc.gov/research/idb). We collected data on civil cases from our seven district courts and appeals in the Ninth Circuit for our timeframe.

We subsetted to civil rights cases by dropping all cases that do not have a Nature of Suit code beginning with 44.[1] Next, we consolidated the 34 possible case outcomes coded by FJC into seven categories: settlements, voluntary dismissals, involuntary dismissals (called "dismissals for other reasons" in the IDB), judgments for the plaintiff, judgments for the defendant, judgments for other (these were judgments for both parties or where winner was missing in the IDB), and all other outcomes. We drop all cases in our dataset that were not yet terminated.

Step 2. Docket Sheets With the cleaned FJC IDB dataset in hand, we then iterate over our docket sheets to supplement with additional information. We link each docket sheet to an entry in the cleaned IDB using the docket numbers available in both data sources. From the docket sheets, we extract the name of the presiding judge, the names of each litigant and their attorneys.

With the litigant information we extracted from the docket sheets, we first classified each party to each case as one of several types: individual, business, government, law enforcement or other. We used machine learning classification to do this. We had a research assistant hand code a random sample of 3,632 parties from our dataset into many categories which we then consolidated into the five categories we list above. We then trained a classifier on this hand-coded sample using the text of the party names to predict categories. We then used these classification models to categorize each party in our entire dataset.

[1] While others have noted that Nature of Suit codes may be inaccurate, this does not pose a problem for us as long as coding errors are uncorrelated with the judges assigned to cases. Indeed, they are typically chosen by the court or attorneys before assignment to a judge. At worst, we believe Nature of Suit coding errors will simply introduce random noise into our estimates.

Using all of the information we collected on litigants (plus the categories we coded as described above), we generated case level counts of the following:

- the number of plaintiffs, defendants and other litigants,[2]
- the number of pro se[3] plaintiffs, pro se defendants and pro se others,
- number of plaintiff attorneys, defendant attorneys and other attorneys,
- the number of plaintiffs, defendants, others and attorneys in each case that were "repeat players" in our dataset (i.e., appeared more than twice), and
- the number of plaintiffs, defendants, other litigants and attorneys on each case that were in each of the five categories listed above (individual, business, etc.).

Step 3. Judges For our causal identification assumption, we must determine which judge was (as-if randomly) assigned to each case at case filing. As we discuss in the main text, judges may influence the trajectory of a case and can only be considered exogenous to case characteristics when they are randomly assigned at case filing.

Unfortunately, the FJC does not report judge names on its publicly released datasets, including the IDB.[4] We extract this information directly from the docket sheets. Each docket sheet contains a field `ASSIGNED TO` which lists the judge assigned to the case. However, this field lists the *last* assigned judge, not the judge assigned to the case at filing. Because the process by which cases get reassigned may be judge-dependent, this may violate our causal identification assumption that cases are as-if randomly assigned to judges.

[2] We reclassified ant litigant called a "petitioner" as a "plaintiff" and any litigant called a "respondant" as a "defendant."

[3] "Pro se" is the legal term for litigants who have no attorney and self-represent.

[4] We cannot clearly discern why the FJC omits this data. Doing so has made it more difficult for researchers to study and better understand how the federal courts work.

We account for this by using automated methods to scan the docket entries to see (1) whether the case was ever reassigned after the filing date, and (2) if other judges were named in the docket entries before this reassignment. Upon identifying the judge assigned at filing, we link their names to the FJC's Biographical Directory of Article III Federal Judges.[5] We link names to this Directory using a custom python package (available on request). This dataset gives us characteristics for each judge, including the party of their appointing president and their status at the time of the case filing (e.g., whether they are a senior judge, chief judge, or visiting judge).

We now describe the process in more detail. To keep the description clear, we refer to the judge listed in the `ASSIGNED TO` field as the "Listed Judge" and the judge we extracted from the docket entries as serving on the case at time of filing as the "First Judge." The problem we seek to address is that the Listed Judge may not be the First Judge.

It is relatively straight forward to extract the name of the Listed Judge from each docket sheet. After doing so, we drop those Listed Judges who were (1) magistrates, (2) not appointed as Article III judges within 30 days of the filing date. In the case of the latter, we assume that these judges were assigned to the case at a date after filing.

We next determine whether there was a First Judge that differs from the Listed Judge. To do this, we scan the docket entries from beginning of the case until the first mention that the case is assigned or reassigned to another judge. We determine this by searching for the terms "assign" or "reassign" and "judge." During our scan, we extract the name of any Article III judge we find. The list of judges we extract from this process is chronological. From this list we then drop any judge who is a chief judge or Listed Judge. We drop chief judges since they are often administratively responsible for reassigning cases, and we wish to capture the name of the judge from whom the case is reassigned, not the judge performing the reassignment. The first judge of those remaining becomes case's First Judge. Note that this process yields the First Judge to take an action in a case (e.g., scheduling an initial hearing), so our presumption is that such a judge, if different from the

[5] Available at https://www.fjc.gov/history/judges.

Listed Judge, was the one who was initially assigned the case.

This process yields at most two judges: a Listed Judge and a First Judge. Either or both of these could be blank for a particular case. For example, if the Listed Judge on a case was a magistrate judge, we drop it and that cell in our dataset is blank for that case. Or, if we do not identify any First Judge for a case using the docket scanning process above, that cell in our dataset would be blank for that case.

To briefly summarize, we begin with 97,725 civil rights cases. For 11% of these cases, we have no Listed Judge. For the majority of these cases, we have no Listed Judge because a magistrate was assigned to the case.[6] For 14% of these cases, we identify a First Judge using the procedure we describe above. For our main analyses, we drop all cases where we identify a First Judge that is different than the Listed Judge.

A.3 Cleaning the Data

Before proceeding with our main analyses, we perform some additional cleaning of the dataset.

First we drop all cases for which we cannot identify an Article III judge as the presiding judge. This amounts to 11% of the dataset, as described above. Then we drop cases that relate to denials of *in forma pauperis* status since we suspect these cases do not follow the typical procedures. This accounts for 0.6% of our dataset. Finally, for our main analysis, we drop all cases where we identify a First Judge that differs from the Listed Judge. This amounts to 14% of our dataset. (In Table C.1, we also present results if we do not drop this latter set of cases and impute the First Judges as the judges assigned to cases, if one is identified.)

As we describe in the text, our causal identification assumption is that cases are block randomized within court divisions and years. As a result, for each analysis, we drop all cases in division-year blocks that have no variation on the treatment (cases all heard by Republican appointees or

[6] As far as we can tell, both the Northern District of California and the District of Oregon allow (or used to allow) some cases to be directly assigned to magistrate judges without also being assigned to a supervising district judge.

vice versa). For our main analysis, the entire process we describe in this section yields a dataset of 70,680 cases, which comprises 72% of the total number of civil rights cases heard by these seven districts in the time frame of our study.

B Using Machine Learning for Causal Inference

In Section 3.1, we describe a novel procedure that provides strong evidence for our causal identification assumption that cases are as-if randomly assigned to judges within court-division-year blocks. In this section, we provide additional motivation for and details about our novel method.

B.1 Weaknesses of Standard Balance Tests

It is standard to provide empirical evidence that a causal identification assumption is justified. The basic idea is to see whether observed pre-treatment variables are predictive of treatment. If they are, then one has detected a threat to causal inference. In our context, a standard approach would be to use an F-test or chi-square test to see whether the observed pre-treatment variables predict treatment above and beyond a baseline model accounting for the division-year blocks. However, these standard tests have at least three weaknesses.

First, there is evidence that these tests can falsely reject the null hypothesis at high rates (for example, see Gerber and Green 2005; Lee 2013). These false positives can induce a researcher to take steps to correct for bias that does not exist. Some of these corrections may even introduce bias (Greenland, Pearl, and Robins 1999).

Second, the threshold of "statistical significance" used by these tests says little about the substantive importance of any threats to causal identification. In observational settings, it will rarely be the case that an assumption of (conditional) random assignment is actually true.[7] With a large

[7] For example, while we assume that cases are randomized to judges within divisions and years, randomization does not actually occur on a yearly basis.

enough dataset, one would expect to detect some imbalance on pre-treatment covariates. These "failures" of randomization could therefore be a result of large sample sizes combined with arbitrary testing thresholds (e.g., p-values less than 0.05). Importantly, an imbalance may be so small in magnitude that it would not appreciably bias estimates.[8]

Third, each of these tests require an analyst to specify an empirical model. It may be possible to "pass" a balance test with some specifications, but not others. But since it is impossible to know the form of potential bias, we do not know which specification is the right one to use when conducting these conventional balance tests. Some model specifications will perform better than others at predicting treatment using pre-treatment variables. Perhaps a model with first-order interactions would predict treatment better than one without? Or perhaps a model with some combination of main terms, first-order interactions, and second-order interactions? Applied researchers rarely explore a wide range of specifications. For example, we conducted numerous F-tests and chi-square tests and estimated a wide range of test statistics, depending on the specification we used.

B.2 Why Machine Learning?

We conduct a balance test that is much more principled and aggressive (at finding imbalances) than standard approaches. We describe the basic process in the main text and refer readers to Section 3.1 for that description. Our procedure depends on obtaining high quality estimates of the probability that each case will be assigned to treatment (i.e., a Republican appointee) or control (a Democratic appointee). That is, we need to estimate "propensity scores" for each case. Rather that use a standard technique for estimating propensity scores, such as logistic regression, we use machine learning. In the remainder of section, we briefly describe the rationale for, and the implementation of, our machine learning processes. Some of the discussion here is technical. For a brief overview of

[8] For example, in one of our many efforts to test for covariate imbalance, we tried to predict treatment with two regressions of pre-treatment variables. The first only used division and year as predictors and the other included division and year plus over 100 pre-treatment variables. The latter model barely nudged the R-squared (from 0.1560 to 0.1596), but a standard F-test yielded a p-value of less than 0.01.

machine learning techniques as used in legal applications, see Copus, Hübert, and Laqueur (2019).

As in the main text, let R_{idy} be a dummy variable that indicates whether case i in division-year dy is assigned to a Republican appointee. Then, any *prediction* of R_{idy} using a statistical model (such as an OLS regression) is a propensity score for case i. We denote this prediction as $\hat{p}_{idy}(X_{idy}) \equiv \widehat{\Pr}(R_{idy} = 1)$, which is the propensity score for case i that is estimated using (pre-treatment) variables X_{idy}. To reduce the notation in what follows, we will drop the dy subscripts. We will use boldface to denote vectors or matrices.

We seek the most accurate predictions possible. That is, we seek to estimate propensity scores $\hat{p}_i(X_i)$ that are as close to the "true" probability of treatment as possible. Accuracy is important. The more inaccurate our predictions, the less aggressive our balance test. And since we are *trying* to detect violations of randomization, inaccurate predictions may cause us to falsely conclude that the conditional independence assumption is satisfied when it is not.

We use machine learning to generate these predictions since machine learning algorithms are typically optimized to produce the most accurate predictions. To assess the accuracy of our predictive models, we will use a standard metric for models with binary dependent variables, the "area under the curve." In most applications, this metric varies from 0.5 to 1.[9] An AUC of 0.5 means that an estimated model produces predictions that are completely inaccurate. That is, the predictions are effectively generated randomly and thus not at all predictive of the actual outcomes. An AUC of 1 means that an estimated model produces predictions that are perfectly accurate. Formally, the AUC is measuring the area under the receiver operating characteristic (ROC) curve, which characterizes the extent to which the estimated model makes type 1 errors (false positives) when generating predictions.[10] Graphically, an ROC curve will lie on the 45-degree line if the estimated

[9] It is possible to have AUC less than 0.5 but this is not typical for most machine learning applications, so we do not discuss it.

[10] Specifically, it plots the true positive rate against the false positive rate when using different thresholds to classify observations. If the model produces accurate predictions independent of the threshold used, then it is perfectly accurate (AUC = 1). If it produces predictions that are equally likely to be true positives or false positives regardless of the threshold, then it is completely inaccurate (AUC = 0.5).

model is completely inaccurate (AUC is 0.5) and will protrude to the north west away from the 45-degree line as the accuracy of the estimated model increases.

Practically speaking, to maximize the accuracy of our predictions, we use a stacked ensemble approach that combines information from several models at once (details below). This approach is known to generate weakly more accurate predictions than if one were to estimate a single model (van der Laan, Polley, and Hubbard 2007).

One problematic way to maximize the accuracy of predictions is to *overfit*. Roughly speaking, overfitting involves estimating a model that generates artificially accurate predictions in the sense that the model generates very accurate predictions for the data used to estimate the model but would not generate accurate predictions in other datasets. To prevent overfitting, all of our predictions are generated from models that were estimated ("trained" in the language of machine learning) on other data. For example, if we want to generate a prediction for a case i, we will generate that prediction from a model that was estimated on a dataset that did not include case i. This is often referred to as "out of sample prediction." More specifically, we accomplish this out of sample prediction using cross-validation—a process where we set aside some of the data, estimate a model on the rest of the data and then predict in the subset of data we initially set aside.

We need to generate predictions $\hat{p}_i(X_i)$ for our propensity scores. So, our goal is to use machine learning to estimate models that yield the most accurate predictions possible without overfitting. For each model, we used a fairly standardized process. We describe each step of the process below.

Step 1: truncate outliers. We begin with some standard data cleaning. Since some algorithms are known to perform worse in the presence of outliers, we truncate many of the predictor variables in X_i. We label this cleaned dataset as X_i'

Step 2: estimate base models. We estimate three "base models" using standard algorithms available in the `h2o` package for python: a Random Forest, a LASSO regression, and an OLS regression. In order to avoid overfitting, we estimate each of these models using 10-fold cross validation with randomly generated folds. After estimating the models, we generate six sets of predictions, one from each of the base models. Note that all predictions are cross-validation predictions and *not* in sample predictions.

Step 3: estimate an ensemble model. Using the six sets of cross validation predictions generated from our base models, we estimate an ensemble model. To do this, we regress R_i onto $\hat{\mathbf{p}}$, where $\hat{\mathbf{p}}$ is a design matrix consisting of the six sets of predictions from our base models. This regression yields a model with predictions that are generated from a weighted average of the predictions from our base models. (We constrain the regression to have non-negative weights.) We cross-validate this regression to generate out-of-sample predictions.

The predictions generated from the ensemble model in Step 3 are our propensity scores. Note that AUC is always weakly higher for the ensemble model than for any of the base models that are used to create the ensemble model.

B.3 The Predictors

We generate our propensity scores using a set of K pre-treatment variables as predictors for our machine learning models. We now briefly describe each variable.

As described in Section 2, the pre-treatment variables are extracted from two sources, the FJC's IDB and our docket sheet collection. We briefly list and describe each of the pre-treatment variables we use for our machine learning models. The following variables come from the IDB:[11]

- `fe`: The court division and year in which the case was filed (the "blocks" we use in our main analyses)
- `nature_of_suit`: The case's nature of suit code
- `SECTION`: The section of the U.S. Code under which the lawsuit was filed
- `ORIGIN`: The case's origin
- `JURIS`: The basis for the court's jurisdiction

[11] More details on these variables available in the IDB codebook at https://www.fjc.gov/sites/default/files/idb/codebooks/Civil%20Codebook%201988%20Forward_0.pdf

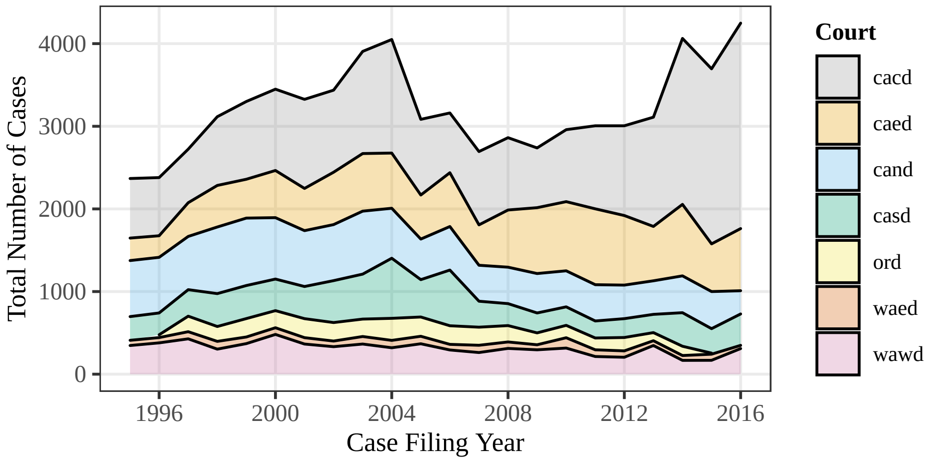
- `jury_demand`: The party or parties that demanded a jury
- `PROSE`: The party or parties that were pro se
- `COUNTY`: The county of residence for the first listed plaintiff

The following variables come from the docket sheets:

- `[pla/def/oth]_count`: The number of [plaintiffs/defendants/other litigants]
- `[pla/def/oth]_count_prose`: The number of pro se [plaintiffs/defendants/other litigants]
- `[pla/def/oth]_count_anonymous`: The number of anonymous [plaintiffs/defendants/other litigants]
- `[pla/def/oth]_count_IND`: The predicted number of [plaintiffs/defendants/other litigants] that are individuals
- `[pla/def/oth]_count_BUS`: The predicted number of [plaintiffs/defendants/other litigants] that are businesses
- `[pla/def/oth]_count_GOV`: The predicted number of [plaintiffs/defendants/other litigants] that are governments
- `[pla/def/oth]_count_LAW`: The predicted number of [plaintiffs/defendants/other litigants] that are related to law enforcement
- `[pla/def/oth]_count_OTH`: The predicted number of [plaintiffs/defendants/other litigants] that were not clearly classified as a type
- `[pla/def/oth]_count_repeat`: The number of [plaintiffs/defendants/other litigants] on this case that appeared in at least one other case in our dataset
- `[pla/def/oth]_acount`: The number of attorneys representing the [plaintiffs/defendants/other litigants]
- `[pla/def/oth]_acount_repeat`: The number of attorneys representing the [plaintiffs/defendants/other litigants] on this case that appeared in at least one other case in our dataset

Except for the court division and year blocks (`fe`), these pre-treatment variables do not play a primary role in our analysis. They are only used to estimate propensity scores. Since most of these variables are effectively auxiliary variables, we do not describe them here. (Data will be made publicly available upon publication.) However, since the court division and year blocks do play a primary role in our analysis (by way of our identification strategy), Figure B.1 plots the number of cases that come from each court in year.

Figure B.1: *We plot the total number of civil rights cases filed in the courts in our dataset. Note that for the District of Oregon, our data spans from mid-1996 through 2015.*



C Analysis: Full Dataset

In this section, we present the numerical estimates of our main effects, as well as some additional details and robustness checks.

C.1 Numerical Estimates

In Table C.1, we provide numerical results corresponding to our main effects, as well as four additional analyses, which we describe in the caption of the table.

Table C.1: *This table presents ARA effects derived from equation (1) in the main text. We present five sets of results. Column (1) presents the results for our main analysis. Column (2) presents results after dropping cases heard by senior judges. For column (3), we present results on a larger dataset where we try to infer the First Judge if she/he is different from the Listed Judge (see Step 3 in Section A.2). Column (4) presents results for all cases heard between 1996 and 2012 for our comparison to the Ninth Circuit (see Section 4.1). Column (5) presents our main results controlling for the assigned judge's years of service as a district judge (see Section 3.4 and Appendix C.3.2). Judge-clustered robust standard errors are reported in parentheses.*

	(1)	(2)	(3)	(4)	(5)
			Pro-Defendant Outcomes		
ARA Effect	0.050	0.064	0.043	0.045	0.049
	(0.010)	(0.011)	(0.009)	(0.011)	(0.010)
			Settlements		
ARA Effect	–0.049	–0.057	–0.040	–0.044	–0.049
	(0.010)	(0.011)	(0.009)	(0.011)	(0.010)
			Involuntary Dismissals		
ARA Effect	0.031	0.034	0.026	0.018	0.031
	(0.008)	(0.008)	(0.007)	(0.008)	(0.008)
			Voluntary Dismissals		
ARA Effect	0.014	0.019	0.014	0.020	0.014
	(0.008)	(0.009)	(0.007)	(0.009)	(0.009)
			Judgments for Defendant		
ARA Effect	0.005	0.011	0.003	0.007	0.005
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
			Judgments for Other		
ARA Effect	–0.003	–0.005	–0.001	0.000	–0.003
	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)
			Judgments for Plaintiff		
ARA Effect	–0.001	–0.002	–0.001	0.000	–0.001
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
			Other Outcomes		
ARA Effect	0.003	0.000	–0.001	–0.001	0.003
	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Observations	70680	60383	85297	53196	70680
Treatment Judges	95	79	134	90	95
Control Judges	98	83	135	80	98
Division-Years	290	264	311	229	290

C.2 Biasing the Estimates

In Section 3.3, we demonstrate that subsetting based on outcomes (which much of the prior research has done) would bias estimates downward. Here, we present a table of numerical estimates corresponding to Figure 5.

Table C.2: *This table presents ARA effects derived from equation (1) in the main text, using prodefendant as our outcome variable. We present five sets of results. Column (1) presents the unbiased, causal results for our main analysis. Column (2) presents biased results after dropping cases that were settled. Column (3) presents biased results after dropping cases that were settled or voluntarily dismissed (i.e., not withdrawn). Column (4) presents biased results for the subset of cases that ended with a formal judgment. Column (5) presents biased results for the subset of district court cases that were eventually appealed to the Ninth Circuit. Judge-clustered robust standard errors are reported in parentheses.*

	(1)	(2)	(3)	(4)	(5)
	Pro-Defendant Outcomes				
ARA Effect	0.050	0.015	0.021	0.015	0.034
	(0.010)	(0.008)	(0.010)	(0.014)	(0.011)
Observations	70680	46305	32850	13659	7577
Treatment Judges	95	93	93	91	87
Control Judges	98	98	96	91	89
Division-Years	290	284	282	277	272

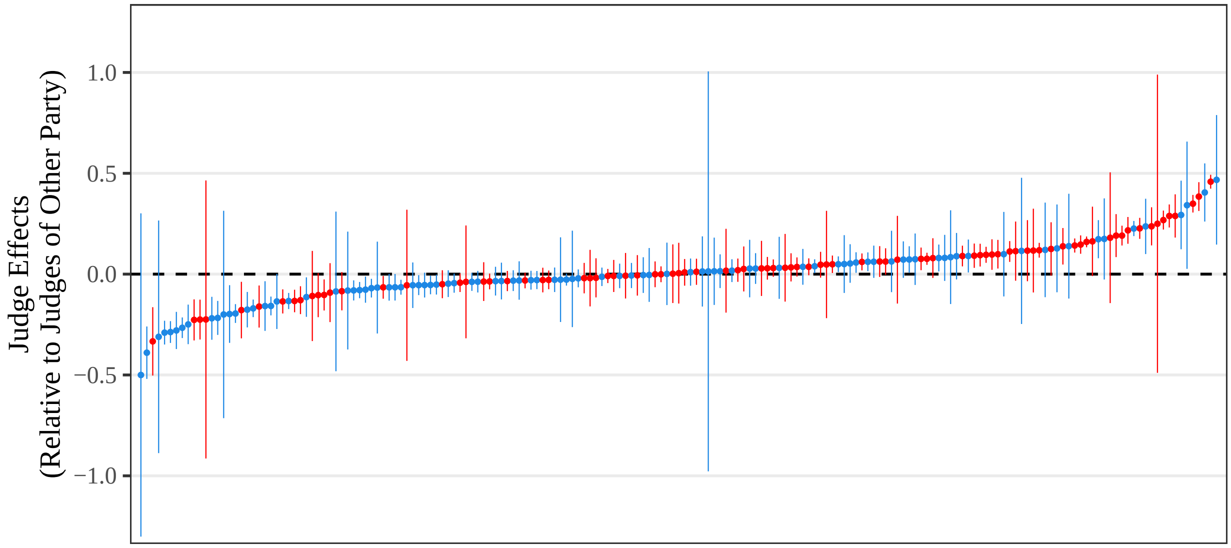
C.3 Robustness Checks

We now provide additional details about the three robustness checks we discuss in Section 3.4.

C.3.1 Judge-Specific Effects

We first check that our main ARA effects are not driven by outlier appointees. To do so, we estimate "judge-specific" effects. Specifically, for each judge, we estimate an effect assuming cases heard by the specific judge are the "treatment group" and cases heard by all judges appointed by presidents of the *other party* are the "control group." We plot the results in Figure C.2. While there is heterogeneity across judges, it is also apparent that the effects are smoothly distributed across judges.

Figure C.2: *We estimate judge-specific effects. For each judge, we estimate the treatment effect of assignment to that judge relative to assignment to another judge appointed by a president of the opposite party. Blue dots indicate judge-specific effects for judges appointed by Democratic presidents whereas red dots indicate judge-specific effects for judges appointed by Republican presidents. Robust standard errors are reported for each judge-specific effect.*



C.3.2 Controlling for Time on Bench

In column (5) of Table C.1, we report the results of our main regression in equation (1), but adding an additional variable for the number of years the assigned judge had been serving as a district judge at the time of the case's filing. Even with this control variable added, the ARA effects are nearly identical.

C.3.3 Effects for Adjacent Presidents

To estimate effects by appointing presidents, we conduct separate sub-analyses for each pair of successive presidents of differing parties. We do separate analyses because some of the division-year combinations do not include cases assigned to each of the presidents' appointees. For each analysis, we include only division-years where both presidents had appointees who were assigned cases.[12] We use the same specification we used in our estimation of ARA effects in Section 3.2.

Figure C.3: *For each pair of adjacent presidents from different parties, we plot the ARA effect with a solid black dot and its corresponding 95% confidence interval (based on judge clustered standard errors). A weighted average of these effects is 0.069.*

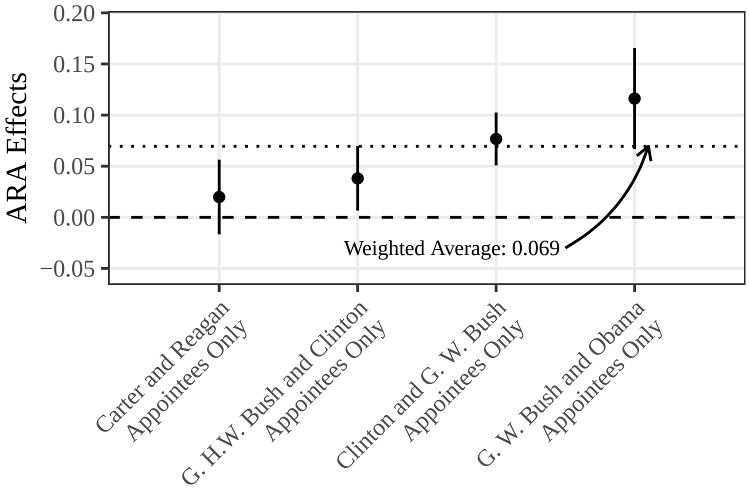


Figure C.3 displays the results of our adjacent president analysis. Note that the point estimates are in the expected direction: we estimate that assignment to a Republican president's appointee rather than a Democratic president's appointee increases the probability of a pro-defendant outcome for each pair of presidents, though the effect for Carter vs. Reagan is not statistically significant. A weighted average of the successor president estimates is statistically indistinguishable from our main ARA effect on pro-defendant outcomes, providing confidence that ARA effects are indeed driven by differences in the types of judges that Republicans and Democrats appoint, and not unrelated time trends.

We present corresponding numerical results in Table C.3.

[12] We exclude appointees of Kennedy, Johnson, Nixon, and Ford due to a low number of observations. We do not analyze a subset of Reagan and G. H.W. Bush appointees since both presidents are from the same party.

Table C.3: *This table presents ARA effects derived from equation (1) in the main text. We present four sets of results. Column (1) presents ARA effects for cases heard by appointees of Carter and Reagan only. Column (2) presents ARA effects for cases heard by appointees of G. H.W. Bush and Clinton only. Column (3) presents ARA effects for cases heard by appointees of Clinton and G. W. Bush only. Column (4) presents ARA effects for cases heard by appointees of G. W. Bush and Obama only. Judge-clustered robust standard errors are reported in parentheses.*

	(1)	(2)	(3)	(4)
	Pro-Defendant Outcomes			
ARA Effect	0.020	0.038	0.077	0.116
	(0.015)	(0.014)	(0.013)	(0.023)
Observations	9875	25909	28689	13160
Treatment Judges	25	38	34	30
Control Judges	17	19	37	25
Division-Years	94	174	156	64

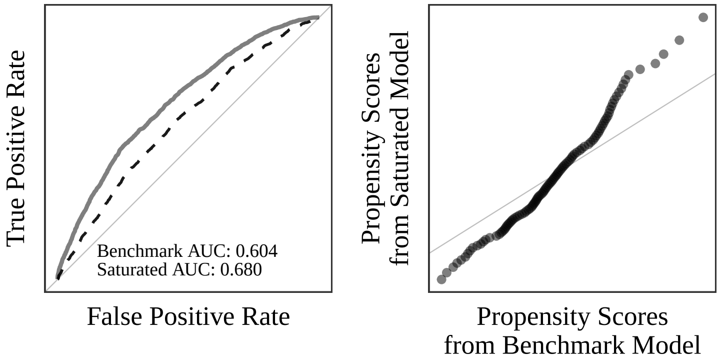
D Ninth Circuit Analysis

We have data on civil rights cases appealed to the Ninth Circuit from 1996 to 2012. We collected this data using a process similar to the one described in Section A. We use this data to calculate the effect on case outcomes of assignment to a majority Republican panel versus a majority Democratic panel.

Our dataset contains 7,654 civil rights appeals heard by three judge panels. Of these appeals, 64% were resolved with a pro-defendant outcome (i.e., reversing a district court decision that favors the defendant or affirming a district court decision that favors the plaintiff).

As we discuss in the main text, we find evidence that cases are not randomly assigned to panels in our dataset of Ninth Circuit appeals heard by three-judge panels. We use the procedure we outline in Section 3.1 on our Ninth Circuit dataset and we plot the corresponding ROC curve and eQQ plot in Figure D.4.

Figure D.4: *Using the same method we describe in Section 3.1, we find evidence that cases are not randomly assigned to panels in our sample of all civil rights appeals filed in the Ninth Circuit and heard by three judge panels from 1996 to 2012. We plot an ROC curve and an eQQ plot to demonstrate this threat to causal inference.*



Since we find evidence that cases are not randomly assigned, we reported inverse probability weighted estimates (IPW) and augmented inverse probability weighted estimates (AIPW) in the main text. Our preferred estimates are the AIPW estimates since they are doubly robust to model misspecification.

We calculate our estimates following Glynn and Quinn (2010). Let i index cases and let $\hat{p}_i(X_i)$ be the case-specific propensity score, which we estimate using the machine learning technique described in Section B. Then, our IPW estimate is calculated as follows:

$$\widehat{\text{IPW}} = \frac{1}{\sum_i \frac{R_i}{\hat{p}_i(X_i)}} \left[\sum_i \frac{R_i Y_i}{\hat{p}_i(X_i)} \right] - \frac{1}{\sum_i \frac{(1-R_i)}{(1-\hat{p}_i(X_i))}} \left[\sum_i \frac{(1-R_i)Y_i}{(1-\hat{p}_i(X_i))} \right]$$

Note, as described in Glynn and Quinn (2010), we renormalize the IPW weights so that they sum to one.

Let $\hat{Y}_i(X_i, R_i) \equiv \hat{Y}_i^{R_i}$ be the predicted outcome of case i based on pre-treatment covariates X_i for treatment group R_i. We estimate these quantities using the machine learning the same techniques described in Section B.

Then, we calculate the AIPW estimate as follows:

$$\widehat{\text{AIPW}} = \frac{1}{N}\sum_i \left[\frac{R_i Y_i}{\hat{p}_i(X_i)} - \frac{(1-R_i)Y_i}{(1-\hat{p}_i(X_i))} - \frac{R_i - \hat{p}_i(X_i)}{\hat{p}_i(X_i)(1-\hat{p}_i(X_i))}\left((1-\hat{p}_i(X_i))\hat{Y}_i^1 + \hat{p}_i(X_i)\hat{Y}_i^0\right)\right]$$

Finally, for each of these estimates, we calculate bootstrapped standard errors clustering at the judge-level. We use the procedure described in section 3.5.2 of Aronow and Miller (2019). Because our clusters are unequal sizes and we have many cluster sizes, we perform an adjustment described in Sherman and Cessie (1997, p. 905). Specifically, we proceed assuming equal cluster sizes, and then weight each bootstrap estimate by $\sqrt{N^*/N}$ where N^* is the sample size of the bootstrap sample and N is the sample size of our entire dataset. Note that because each bootstrap iteration randomly samples *clusters* (which are differing sizes), N^* will be generically unequal to N.

Table D.4: *We present the results of our analysis of civil rights appeals in the Ninth Circuit. We present three sets of results: unadjusted estimates, inverse probability weighted estimates, and augmented inverse probability weighted estimates. We bootstrap panel clustered standard errors. We describe our estimation procedure in more detail in the text.*

	Unadjusted Estimate	IPW Estimate	AIPW Estimate
Treatment Effect	0.0777	0.0603	0.0538
	(0.0140)	(0.0149)	(0.0098)
Observations	7,862	7,862	7,862
Majority Republican Panels	929	929	929
Majority Democratic Panels	1,427	1,427	1,427
Division-Years	51	51	51

In Table D.4, we report the results of our analysis of Ninth Circuit cases.

E ARA Effects Over Time

In the main text, we explore how ARA effects change over time. In this section, we present numerical results corresponding to Figure 7.

Table E.5: *This table presents ARA effects derived from equation (1) in the main text. We present three sets of results. Column (1) presents ARA effects for cases filed 1995–2000. Column (2) presents ARA effects for cases filed 2001–2008. Column (3) presents ARA effects for cases filed 2009–2016. Judge-clustered robust standard errors are reported in parentheses.*

All Appointees

	(1)	(2)	(3)
		Pro-Defendant Outcomes	
ARA Effect	0.010	0.052	0.074
	(0.015)	(0.015)	(0.016)
Observations	17335	26520	26825
Treatment Judges	58	75	53
Control Judges	61	52	76
Division-Years	67	114	109

Reagan and Clinton Appointees

	(1)	(2)	(3)
		Pro-Defendant Outcomes	
ARA Effect	0.052	0.057	0.130
	(0.011)	(0.015)	(0.026)
Observations	9147	10355	4166
Treatment Judges	27	18	6
Control Judges	37	34	15
Division-Years	52	67	33

G. H.W. Bush and Clinton Appointees

	(1)	(2)	(3)
		Pro-Defendant Outcomes	
ARA Effect	0.025	0.036	0.074
	(0.010)	(0.016)	(0.034)
Observations	8221	13970	3718
Treatment Judges	17	17	14
Control Judges	35	35	21
Division-Years	51	83	40

G. W. Bush and Clinton Appointees

	(1)	(2)	(3)
		Pro-Defendant Outcomes	
ARA Effect	—	0.068	0.084
	—	(0.011)	(0.020)
Observations	0	13516	15173
Treatment Judges	0	33	32
Control Judges	0	35	33
Division-Years	0	70	86

Obama and Clinton Appointees

	(1)	(2)	(3)
		Pro-Defendant Outcomes	
ARA Effect	—	—	–0.004
	—	—	(0.028)
Observations	0	0	9445
Treatment Judges	0	0	33
Control Judges	0	0	32
Division-Years	0	0	64

F A Back of the Envelope Calculation: Trump's Impact

In the conclusion, we present a back of the envelope calculation of the overall impact on civil rights cases of Donald Trump's appointments to the district courts. We now provide some additional details about this calculation.

First, using FJC's Biographical Directory of Article III Federal Judges appointed from Johnson to Reagan,[13] we estimate that, on average, an appointee will serve 15 years as an active judge and nine years as a senior judge.

Second, using statistics from the FJC's Integrated Database, we estimate that since 1990 there has been an average of approximately 37,500 civil rights cases per year across all district courts in the U.S.

Third, using information provided by the federal courts,[14] we estimate that 15% of the federal caseload is heard by senior judges.

Fourth, using FJC's Biographical Directory of Article III Federal Judges, we estimate that during the most recent years (2008-2016), the average number of active judges on bench is 618 while the average number of senior judges on the bench is 419.

Fifth, we assume that Donald Trump will appoint 150 district court judges during his first term (as of July 30, 2020, he has appointed 147) and that Hillary Clinton would also have appointed 150 district court judges.

We assume that we can project our treatment effects into Trump appointees (and counterfactual Hillary Clinton appointees) and that case outcomes are only affected through district court appointments.[15] Then, we calculate the back of the envelope estimate as follows. First, using the

[13] See `https://www.fjc.gov/history/judges`. We do not include appointees prior to Johnson since life expectancy has increased, as well as norms and policies around senior service. We do not include appointees after Reagan since many of the more modern appointees have not finished their active or senior service.

[14] See `https://www.uscourts.gov/faqs-federal-judges#faq-What-is-a-senior-judge?`.

[15] Of course, presidential elections and appointments may and probably do affect civil rights in ways other than through observable case outcomes. For example, new judicial appointments might alter people's assessments of their chances of winning, causing them to file different types of cases. Or companies might also change their business

benchmark statistics above, we estimate the total number of civil rights cases that a typical district judge hears during their entire time serving as a district judge:

$$\underbrace{\frac{37,500 \times 0.85}{618} \times 15}_{\text{total cases heard as an active judge}} + \underbrace{\frac{37,500 \times 0.15}{419} \times 9}_{\text{total cases heard as a senior judge}} \approx 900 \text{ civil rights cases during typical life tenure}$$

Second, since we assume that Trump will (and Clinton would have) appointed 150 district court judges, we estimate that approximately 900 × 150 = 135,000 civil rights cases will be heard by Trump appointees, and that the same number would have been heard by Clinton appointees.

Third, if we assume that we can project our main ARA effect into Trump appointments, then 135,000 × 0.05 ≈ 6,750 cases will be resolved in favor of defendants because of Trump's election and appointment of new district judges. And, if we assume that we can project our most recent ARA effect (for cases heard from 2008-2016) into Trump appointees, then 135,000 × 0.074 ≈ 9,990 cases will be resolved in favor of defendants because of Trump's election and appointment of new district judges rather than Clinton's.

References

Aronow, Peter M., and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics.* New York: Cambridge University Press.

Copus, Ryan, Ryan Hübert, and Hannah Laqueur. 2019. "Credible Prediction: Machine Learning and the Credibility Revolution." In *Law as Data: Computation and the Future of Legal Analysis.* SFI Press.

Gerber, Alan S., and Donald P. Green. 2005. "Correction to Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005)." *American Political Science Review* 99 (2): 301–313.

Glynn, Adam N., and Kevin M. Quinn. 2010. "An Introduction to the Augmented Inverse Propensity Weighted Estimator." *Political Analysis* 18:36–56.

Greenland, Sander, Judea Pearl, and James M. Robins. 1999. "Causal Diagrams for Epidemiologic Research." *Epidemiology* 10 (1): 37–48.

practices in anticipation of new appointees' jurisprudence. We hope future research can address these issues.

Lee, Wang-Sheng. 2013. “Propensity Score Matching and Variations on the Balancing Test.” *Empirical Economics* 44:47–80.

Sherman, Michael, and Saskia le Cessie. 1997. “A Comparison Between Bootstrap Methods and Generalized Estimating Equations for Correlated Outcomes in Generalized Linear Models.” *Communications in Statistics - Simulation and Computation* 26 (3): 901–925.

van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. “Super Learner.” `http://biostats.bepress.com/ucbbiostat/paper222/`.