Supplementary material

A The benefits of pairing DDML and stacking

Table A.1: Mean-squared prediction error

	$n_b = 9\,915$		$n_b = 99\,150$	
Panel (A): Linear DGP	$E[Y\|X]$	$E[D\|X]$	$E[Y\|X]$	$E[D\|X]$
Candidate learners				
OLS	3.093	0.200	3.088	0.200
Lasso with CV (2nd order poly)	3.095	0.200	3.088	0.200
Ridge with CV (2nd order poly)	3.100	0.200	3.089	0.200
Lasso with CV (10th order poly)	3.298	0.202	3.095	0.200
Ridge with CV (10th order poly)	3.423	0.206	3.093	0.200
Random forest (low regularization)	3.613	0.233	3.698	0.239
Random forest (high regularization)	3.183	0.205	3.197	0.207
Gradient boosting (low regularization)	3.131	0.201	3.102	0.200
Gradient boosting (high regularization)	3.151	0.201	3.137	0.201
Neural net	3.227	0.204	3.153	0.200
Panel (B): Non-Linear DGP	$E[Y\|X]$	$E[D\|X]$	$E[Y\|X]$	$E[D\|X]$
Candidate learners				
OLS	3.681	0.203	3.672	0.203
Lasso with CV (2nd order poly)	3.480	0.201	3.449	0.200
Ridge with CV (2nd order poly)	3.479	0.201	3.449	0.200
Lasso with CV (10th order poly)	6.161	0.223	3.422	0.200
Ridge with CV (10th order poly)	7.431	0.230	3.424	0.200
Random forest (low regularization)	3.789	0.231	3.515	0.236
Random forest (high regularization)	3.588	0.204	3.251	0.205
Gradient boosting (low regularization)	3.345	0.200	3.095	0.198
Gradient boosting (high regularization)	3.399	0.200	3.216	0.199
Neural net	3.694	0.205	3.510	0.200

Notes: The table shows the mean-squared prediction error of each candidate learner from the simulation example in Section 4.1. The bootstrap sample size is $n_b = 9\,915$ or $99\,150$. Results are based on 1 000 replications. See Table 1 for more information.

Table A.2: Bias in the Linear and Non-Linear DGP

	Panel (A): Linear DGP				*Panel (B): Non-linear DGP*			
	$n_b = 9\,915$		99 150		$n_b = 9\,915$		99 150	
	Bias	s.e.	Bias	s.e.	Bias	s.e.	Bias	s.e.
Full sample:								
OLS	49.9	39.6	-6.8	12.6	-2588.9	46.5	-2632.3	14.8
PDS-Lasso	48.4	39.6	-4.2	12.6	-2598.7	46.5	-2631.6	14.8
DDML methods:								
Candidate learners								
OLS	46.2	39.7	-6.9	12.6	-2613.0	47.5	-2635.4	14.8
Lasso with CV (2nd order poly)	50.9	39.7	-6.2	12.6	703.7	44.5	718.5	13.5
Ridge with CV (2nd order poly)	48.2	39.9	-6.9	12.6	767.4	44.6	729.3	13.5
Lasso with CV (10th order poly)	248.1	266.6	55.9	12.6	-4109.0	1325.8	7.4	22.5
Ridge with CV (10th order poly)	1230.1	401.8	31.6	12.6	-5126.2	1713.9	9.6	23.3
Random forest (low regularization)	-74.7	48.6	-25.2	16.0	-96.1	48.5	-37.5	15.4
Random forest (high regularization)	69.1	41.7	-23.5	13.2	-159.7	43.5	-4.2	13.0
Gradient boosting (low regularization)	12.1	40.1	-24.2	12.6	8.5	42.7	30.9	12.4
Gradient boosting (high regularization)	114.8	39.8	66.9	12.5	162.0	42.1	200.1	12.6
Neural net	394.2	43.1	9.1	13.6	-601.3	46.5	-131.9	14.3
Stacking approaches								
Stacking: CLS	42.8	40.2	-7.5	12.6	133.9	195.6	37.8	12.4
Stacking: Average	107.7	51.2	-6.5	12.6	94.0	110.3	72.3	12.7
Stacking: OLS	-129.3	96.2	-9.4	14.9	-204.8	405.2	17.5	17.8
Stacking: Single-best	43.7	39.9	-8.6	12.6	-121.9	272.9	30.9	12.4
Short-stacking: CLS	45.0	39.7	-7.0	12.6	162.7	42.0	33.6	12.4
Short-stacking: Average	107.7	51.2	-6.5	12.6	94.0	110.3	72.3	12.7
Short-stacking: OLS	37.6	39.6	-7.8	12.6	123.6	41.7	29.5	12.3
Short-stacking: Single-best	44.4	39.7	-8.3	12.6	71.7	42.6	30.9	12.4
Pooled stacking: CLS	58.6	41.4	-7.1	12.6	209.8	63.7	37.5	12.4
Pooled stacking: Average	107.7	51.2	-6.5	12.6	94.0	110.3	72.3	12.7
Pooled stacking: OLS	46.5	47.8	-7.8	12.6	234.5	124.9	30.6	12.3
Pooled stacking: Single-best	46.9	39.7	-8.3	12.6	103.3	45.8	30.9	12.4

Notes: The table reports mean bias and associated standard errors (s.e.) for the listed estimators from the simulation example in Section 4.1. Results are based on 1 000 replications. See Table 1 for more information.

Table A.3: Average stacking weights using OLS as the final learner

	Stacking		Pooled stacking		Short-stacking	
	$E[Y\|X]$	$E[D\|X]$	$E[Y\|X]$	$E[D\|X]$	$E[Y\|X]$	$E[D\|X]$
Panel (A): Linear DGP and $n_b = 9,915$						
OLS	1.161	0.932	1.014	0.815	0.923	0.686
Lasso with CV (2nd order poly)	0.011	0.034	0.066	0.085	0.154	0.079
Ridge with CV (2nd order poly)	–0.157	–0.187	–0.091	–0.137	–0.080	–0.080
Lasso with CV (10th order poly)	–0.046	0.076	–0.027	0.069	–0.041	0.122
Ridge with CV (10th order poly)	0.	–0.006	0.	0.003	0.017	0.022
Random forest (low regularization)	0.001	–0.005	0.001	–0.006	0.003	–0.006
Random forest (high regularization)	–0.014	–0.006	–0.010	–0.002	–0.011	0.008
Gradient boosting (low regularization)	–0.060	–0.108	–0.045	–0.096	–0.039	–0.094
Gradient boosting (high regularization)	0.096	0.260	0.090	0.259	0.076	0.257
Neural net	–195.245	0.009	0.008	0.034	0.004	0.033
Panel (B): Linear DGP and $n_b = 99,150$						
OLS	1.292	0.935	1.130	0.769	1.008	0.677
Lasso with CV (2nd order poly)	–0.103	–0.497	–0.036	–0.378	0.033	–0.398
Ridge with CV (2nd order poly)	–0.170	–0.053	–0.093	–0.019	–0.044	0.056
Lasso with CV (10th order poly)	0.065	0.407	0.023	0.368	0.009	0.335
Ridge with CV (10th order poly)	–0.071	–0.118	–0.022	–0.068	–0.003	–0.034
Random forest (low regularization)	–0.000	–0.001	–0.000	–0.001	0.	–0.001
Random forest (high regularization)	–0.001	–0.002	–0.000	–0.001	–0.000	–0.001
Gradient boosting (low regularization)	–0.034	0.158	–0.021	0.177	–0.020	0.224
Gradient boosting (high regularization)	0.034	0.072	0.027	0.058	0.022	0.020
Neural net	–12.267	0.110	–0.005	0.108	–0.003	0.133
Panel (C): Non-Linear DGP and $n_b = 9,915$						
OLS	0.013	0.050	0.004	0.038	–0.048	0.022
Lasso with CV (2nd order poly)	–0.118	–0.319	–0.166	–0.212	0.033	–0.347
Ridge with CV (2nd order poly)	0.365	0.528	0.427	0.457	0.146	0.548
Lasso with CV (10th order poly)	–0.014	0.126	0.015	0.118	0.091	0.088
Ridge with CV (10th order poly)	0.060	0.028	0.040	–0.003	–0.011	0.055
Random forest (low regularization)	0.052	–0.016	0.056	–0.017	0.057	–0.016
Random forest (high regularization)	–0.098	0.059	–0.099	0.065	–0.110	0.064
Gradient boosting (low regularization)	1.096	0.029	1.133	0.059	1.358	0.159
Gradient boosting (high regularization)	–0.531	0.526	–0.586	0.508	–0.716	0.426
Neural net	–72.340	0.027	0.121	0.046	0.120	0.054
Panel (D): Non-Linear DGP and $n_b = 99,150$						
OLS	–0.015	0.009	–0.017	0.008	0.	0.009
Lasso with CV (2nd order poly)	0.089	–0.806	0.100	–0.802	–0.026	–0.721
Ridge with CV (2nd order poly)	–0.171	0.792	–0.177	0.789	–0.053	0.696
Lasso with CV (10th order poly)	0.285	–0.247	0.229	–0.243	0.358	–0.325
Ridge with CV (10th order poly)	–0.338	0.320	–0.282	0.315	–0.425	0.383
Random forest (low regularization)	0.142	–0.011	0.142	–0.011	0.168	–0.013
Random forest (high regularization)	–0.111	0.069	–0.108	0.070	–0.125	0.079
Gradient boosting (low regularization)	2.368	1.233	2.356	1.236	2.353	1.332
Gradient boosting (high regularization)	–1.443	–0.403	–1.435	–0.406	–1.424	–0.498
Neural net	–45.762	0.030	0.080	0.033	0.066	0.040

Notes: The table shows the (average) stacking weights of each candidate learner for conventional stacking, pooled stacking and short-stacking using OLS as the final learner from the simulation example in Section 4.1. The bootstrap sample size is denoted by n_b. Results are based on 1 000 replications. See Table 1 for more information.

Table A.4: Average stacking weights using single-best

	Stacking		Pooled stacking		Short-stacking	
	$E[Y\|X]$	$E[D\|X]$	$E[Y\|X]$	$E[D\|X]$	$E[Y\|X]$	$E[D\|X]$
Panel (A): Linear DGP and $n_b = 9,915$						
OLS	0.812	0.645	0.889	0.750	0.821	0.650
Lasso with CV (2nd order poly)	0.163	0.274	0.108	0.234	0.150	0.266
Ridge with CV (2nd order poly)	0.021	0.016	0.003	0.003	0.028	0.027
Lasso with CV (10th order poly)	0.002	0.048	0.	0.011	0.001	0.034
Ridge with CV (10th order poly)	0.002	0.010	0.	0.001	0.	0.023
Random forest (low regularization)	0.	0.	0.	0.	0.	0.
Random forest (high regularization)	0.	0.	0.	0.	0.	0.
Gradient boosting (low regularization)	0.	0.003	0.	0.001	0.	0.
Gradient boosting (high regularization)	0.	0.003	0.	0.	0.	0.
Neural net	0.	0.	0.	0.	0.	0.
Panel (B): Linear DGP and $n_b = 99,150$						
OLS	0.895	0.292	0.960	0.288	0.880	0.181
Lasso with CV (2nd order poly)	0.087	0.093	0.037	0.073	0.078	0.059
Ridge with CV (2nd order poly)	0.018	0.016	0.003	0.003	0.042	0.017
Lasso with CV (10th order poly)	0.	0.186	0.	0.172	0.	0.151
Ridge with CV (10th order poly)	0.	0.410	0.	0.464	0.	0.591
Random forest (low regularization)	0.	0.	0.	0.	0.	0.
Random forest (high regularization)	0.	0.	0.	0.	0.	0.
Gradient boosting (low regularization)	0.	0.002	0.	0.	0.	0.001
Gradient boosting (high regularization)	0.	0.	0.	0.	0.	0.
Neural net	0.	0.	0.	0.	0.	0.
Panel (C): Non-linear DGP and $n_b = 9,915$						
OLS	0.	0.	0.	0.	0.	0.
Lasso with CV (2nd order poly)	0.093	0.149	0.066	0.144	0.054	0.081
Ridge with CV (2nd order poly)	0.132	0.124	0.126	0.106	0.058	0.093
Lasso with CV (10th order poly)	0.072	0.058	0.040	0.028	0.039	0.026
Ridge with CV (10th order poly)	0.019	0.041	0.007	0.015	0.005	0.037
Random forest (low regularization)	0.	0.	0.	0.	0.	0.
Random forest (high regularization)	0.002	0.001	0.	0.	0.002	0.
Gradient boosting (low regularization)	0.673	0.357	0.759	0.403	0.832	0.623
Gradient boosting (high regularization)	0.009	0.268	0.002	0.304	0.008	0.140
Neural net	0.002	0.002	0.	0.	0.002	0.
Panel (D): Non-linear DGP and $n_b = 99,150$						
OLS	0.	0.	0.	0.	0.	0.
Lasso with CV (2nd order poly)	0.	0.	0.	0.	0.	0.
Ridge with CV (2nd order poly)	0.	0.	0.	0.	0.	0.
Lasso with CV (10th order poly)	0.	0.	0.	0.	0.	0.
Ridge with CV (10th order poly)	0.	0.	0.	0.	0.	0.
Random forest (low regularization)	0.	0.	0.	0.	0.	0.
Random forest (high regularization)	0.	0.	0.	0.	0.	0.
Gradient boosting (low regularization)	1.	1.	1.	1.	1.	1.
Gradient boosting (high regularization)	0.	0.	0.	0.	0.	0.
Neural net	0.	0.	0.	0.	0.	0.

Notes: The table shows the (average) rates at which each candidate learner is selected by the single-best final learner when using conventional stacking, pooled stacking and short-stacking. The bootstrap sample size is denoted by n_b. Results are based on 1 000 replications. See Table 1 for more information.

Table A.5: Computational time of DDML with conventional and short-stacking

Folds K	Obs.	DDML Stacking Conv.	DDML Stacking Short	OLS	PDS lasso	Ratio
2	200	14.31	3.81	0.0045	0.0461	0.2661
	400	14.93	4.05	0.0047	0.0492	0.2711
	800	18.36	4.95	0.0050	0.0518	0.2696
	1600	26.82	7.07	0.0056	0.0595	0.2636
	9915	138.39	34.35	0.0116	0.1488	0.2482
	99150	2687.07	620.98	0.1013	1.4431	0.2311
5	200	35.77	8.28	0.0045	0.0573	0.2315
	400	41.83	9.79	0.0047	0.0589	0.2342
	800	56.38	13.47	0.0049	0.0624	0.2388
	1600	91.76	21.34	0.0056	0.0711	0.2326
	9915	589.14	136.35	0.0110	0.1508	0.2314
10	200	72.79	16.19	0.0046	0.0423	0.2224
	400	85.87	19.27	0.0046	0.0524	0.2244
	800	119.48	27.80	0.0049	0.0468	0.2327
	1600	197.85	45.59	0.0054	0.0618	0.2304
	9915	1364.07	313.84	0.0113	0.1426	0.2301

Notes: The table reports the computational time in seconds of DDML paired with conventional stacking ('Conv.') or short-stacking ('Short') as implemented in Ahrens et al. (2024), OLS as implemented in Stata's `regress`, post-double-selection lasso as implemented in `pdslasso` (Ahrens, Hansen, and Schaffer, 2018). DDML uses $V = 5$ cross-validation folds and K cross-fitting folds as indicated. Times reported are in seconds (average over 1 000 replications). The computations were performed on the high-performance cluster of the ETH Zurich. Each instance used a single core of an AMD EPYC processor with 2.25-2.6GHz (nominal)/3.3-3.5 GHz (peak) and 4GB RAM.

B DDML and stacking in very small samples

Table B.1: Estimates based on the full sample ($N = 9\,915$).

Estimator	*Estimate*
Panel A. No sample splitting	
OLS TWI	6751.907
OLS QSI	5988.413
Post double Lasso TWI c=0.5	6562.923
Post double Lasso QSI c=0.5	5648.14
Post double Lasso TWI c=1	6630.751
Post double Lasso QSI c=1	4646.575
Post double Lasso TWI c=1.5	7474.508
Post double Lasso QSI c=1.5	4472.324
Panel B. DDML with candidate learners	
Neural net	6433.092
OLS	6463.73
Lasso with CV (TWI)	6780.161
Ridge with CV (TWI)	6760.134
Lasso with CV (QSI)	5722.624
Ridge with CV (QSI)	5995.346
Random forest (low regularization)	6089.389
Random forest (high regularization)	6552.221
Gradient boosting (low regularization)	7003.373
Gradient boosting (high regularization)	7992.538
Panel C. DDML with stacking approaches	
Neural net	6433.092
OLS	6463.73
Lasso with CV (TWI)	6780.161
Ridge with CV (TWI)	6760.134
Lasso with CV (QSI)	5722.624
Ridge with CV (QSI)	5995.346
Random forest (low regularization)	6089.389
Random forest (high regularization)	6552.221
Gradient boosting (low regularization)	7003.373
Gradient boosting (high regularization)	7992.538

Notes: In the case of DDML estimators, the average estimates and standard errors are based on 50 replications. Panel A is reproduced from Table 1 in WZ.

Table B.2: Short-stacking weights using CLS

Estimator	*Observations*						
	200	400	600	800	1 200	1 600	9 915
Panel A. $E[Y\|X]$, $K = 10$							
OLS	.164	.152	.115	.079	.037	.019	0
Neural net	.047	.045	.048	.067	.098	.05	.076
Lasso with CV (TWI)	.043	.034	.034	.035	.03	.033	.091
Ridge with CV (TWI)	.056	.048	.041	.025	.011	.006	.032
Lasso with CV (QSI)	.252	.274	.266	.264	.271	.297	.639
Ridge with CV (QSI)	.194	.252	.297	.328	.341	.357	.153
Random forest (low regularization)	.095	.097	.113	.131	.161	.2	.01
Random forest (high regularization)	.081	.04	.025	.021	.018	.016	0
Gradient boosting (low regularization)	.041	.04	.049	.041	.03	.021	0
Gradient boosting (high regularization)	.028	.019	.013	.009	.002	.001	0
Panel B. $E[D\|X]$, $K = 10$							
OLS	.132	.196	.234	.252	.245	.257	.163
Neural net	.04	.041	.038	.036	.031	.029	.038
Lasso with CV (TWI)	.053	.031	.025	.02	.016	.012	.106
Ridge with CV (TWI)	.038	.018	.013	.015	.008	.005	.029
Lasso with CV (QSI)	.173	.225	.25	.248	.25	.228	.413
Ridge with CV (QSI)	.202	.124	.072	.06	.068	.064	0
Random forest (low regularization)	.103	.123	.144	.187	.249	.307	.006
Random forest (high regularization)	.159	.129	.107	.09	.051	.031	.102
Gradient boosting (low regularization)	.043	.046	.054	.047	.045	.041	.144
Gradient boosting (high regularization)	.059	.065	.064	.046	.038	.025	0
Panel C. $E[Y\|X]$, $K = 10$							
OLS	.122	.098	.066	.026	.003	.001	0
Neural net	0	0	0	0	0	0	0
Lasso with CV (TWI)	.03	.022	.01	.013	.014	.023	0
Ridge with CV (TWI)	.074	.077	.079	.052	.03	.013	0
Lasso with CV (QSI)	.323	.376	.361	.381	.393	.405	.995
Ridge with CV (QSI)	.239	.314	.379	.428	.478	.479	.005
Random forest (low regularization)	.129	.058	.05	.049	.049	.044	0
Random forest (high regularization)	.022	.005	.001	.001	0	.001	0
Gradient boosting (low regularization)	.025	.033	.046	.046	.032	.034	0
Gradient boosting (high regularization)	.035	.016	.009	.004	0	0	0
Panel D. $E[D\|X]$, $K = 10$							
OLS	.038	.108	.17	.189	.173	.132	.005
Neural net	0	0	0	0	0	0	0
Lasso with CV (TWI)	.058	.032	.017	.011	.005	.003	.002
Ridge with CV (TWI)	.06	.013	.009	.01	.002	.001	0
Lasso with CV (QSI)	.232	.309	.313	.287	.261	.168	.754
Ridge with CV (QSI)	.242	.105	.032	.034	.05	.032	0
Random forest (low regularization)	.079	.028	.011	.008	.004	.003	0
Random forest (high regularization)	.185	.249	.256	.304	.344	.507	0
Gradient boosting (low regularization)	.009	.022	.048	.064	.115	.141	.24
Gradient boosting (high regularization)	.098	.135	.143	.092	.046	.013	0

Notes: The table reports the stacking weights corresponding to the DDML stacking estimator in Figure 3. The stacking weights are averaged over folds, based on 10-fold cross-fitting and shows for the estimation of $E[Y|X]$ and $E[D|X]$ in Panel A and B, respectively. See notes below Table 3 for more information.

Table B.3: Mean bias under linear DGP in small samples based on the calibrated Monte Carlo in Section 4.1

	Bootstrap sample size n_b									
	200		400		800		1 600		9 915	
	Bias	s.e.	Bias	s.e.	Bias	s.e.	Bias	s.e.	Bias	s.e.
Full sample estimators:										
OLS	230.4	272.9	-28.4	198.8	9.7	140.9	11.7	102.0	78.3	39.1
PDS-Lasso	-714.3	270.5	-739.5	197.3	-367.9	141.1	-76.3	102.3	78.8	39.1
DDML methods:										
Candidate learners (K = 10)										
OLS	179.4	275.7	-85.1	199.2	2.5	141.0	7.9	102.0	77.8	39.1
Lasso with CV (2nd order poly)	357.0	271.3	-38.9	199.6	3.4	141.7	9.1	102.1	77.5	39.1
Ridge with CV (2nd order poly)	334.6	267.0	6.4	197.5	10.8	142.0	-9.9	102.0	77.5	39.2
Lasso with CV (10th order poly)	-4864.6	5369.1	754.7	957.7	-476.8	1663.2	-547.1	454.5	141.0	46.4
Ridge with CV (10th order poly)	-14274.3	10353.7	1038.4	7241.9	1596.6	3409.3	1442.5	1128.7	449.8	62.2
Random forest (low regularization)	328.1	302.1	-112.2	220.3	-148.4	153.2	-69.6	112.1	-41.8	45.7
Random forest (high regularization)	1086.7	274.9	255.6	201.6	206.3	143.1	128.5	103.7	68.4	39.9
Gradient boosting (low regularization)	171.1	300.8	-280.4	210.3	-149.1	145.7	-60.3	104.2	44.3	39.3
Gradient boosting (high regularization)	389.4	286.4	-77.4	203.4	65.0	142.9	75.2	102.6	139.1	39.2
Neural net	3887.7	247.4	3926.3	199.2	3504.2	143.1	1874.9	104.9	212.1	39.2
Meta learners (K = 10)										
Stacking: CLS	-1150.3	1159.2	-84.3	224.9	28.2	165.0	47.1	111.1	73.1	39.2
Stacking: Average	-3471.2	3073.6	195.7	358.6	-71.7	357.8	368.6	204.4	86.3	39.3
Stacking: OLS	4.2e6	5.4e6	-1040.5	2688.0	-2085.4	1149.5	-365.2	255.9	32.8	55.8
Stacking: Single-best	-5.0	275.9	-212.2	211.7	-57.8	141.7	-27.6	102.4	68.2	39.1
Short-stacking: CLS	322.3	271.8	-54.8	199.6	51.6	141.2	25.1	102.1	73.2	39.1
Short-stacking: Average	-3471.2	3073.6	195.7	358.6	-71.7	357.8	368.6	204.4	86.3	39.3
Short-stacking: OLS	-214.9	269.4	-278.5	200.0	-100.4	140.4	-41.0	101.6	70.7	39.1
Short-stacking: Single-best	189.1	271.1	-91.4	199.1	0.7	140.8	3.1	101.9	75.9	39.1
Pooled stacking: CLS	465.6	271.7	39.7	199.8	89.6	140.9	18.0	102.1	73.7	39.1
Pooled stacking: Average	-3471.2	3073.6	195.7	358.6	-71.7	357.8	368.6	204.4	86.3	39.3
Pooled stacking: OLS	326.6	272.8	-97.4	200.0	-68.3	141.3	-23.2	102.3	72.6	39.1
Pooled stacking: Single-best	150.8	270.7	-93.1	199.3	-11.6	141.2	8.3	102.1	76.6	39.1
Meta learners (K = 2)										
Stacking: CLS	6329.4	5587.9	-66.9	531.5	-21.5	200.4	261.0	111.9	42.8	40.2
Stacking: Average	-972.6	2751.4	1796.8	1765.4	254.6	647.7	815.2	215.7	107.7	51.2
Stacking: OLS	-10699.3	13002.6	3.8e7	3.8e7	-63.3	2752.6	-1043.6	547.8	-129.3	96.2
Stacking: Single-best	5893.9	5319.3	73.8	249.9	-158.5	149.5	115.1	99.1	43.7	39.9
Short-stacking: CLS	389.3	285.0	314.6	210.6	9.9	144.1	218.4	98.2	45.0	39.7
Short-stacking: Average	-972.6	2751.4	1796.8	1765.4	254.6	647.7	815.2	215.7	107.7	51.2
Short-stacking: OLS	-207.9	275.1	84.7	206.1	-113.2	141.6	124.9	97.9	37.6	39.6
Short-stacking: Single-best	241.8	284.2	209.2	210.2	-37.5	145.1	178.8	98.5	44.4	39.7
Pooled stacking: CLS	-22063.1	22632.2	378.9	322.2	-141.5	166.8	230.4	107.1	58.6	41.4
Pooled stacking: Average	-972.6	2751.4	1796.8	1765.4	254.6	647.7	815.2	215.7	107.7	51.2
Pooled stacking: OLS	-26101.3	29649.6	-2324.4	2415.4	-1641.0	1777.9	-79.0	272.3	46.5	47.8
Pooled stacking: Single-best	12247.4	12403.7	122.1	214.6	-88.2	146.2	156.7	98.6	46.9	39.7

Notes: The table reports mean bias and associated standard errors for the listed estimators. We consider DDML with the following individual learners: OLS with elementary covariates, CV lasso and CV ridge with second-order polynomials and interactions, CV lasso and CV ridge with 10th-order polynomials but no interactions, random forest with low regularization (8 predictors considered at each leaf split, no limit on the number of observations per node, bootstrap sample size of 70%), highly regularized random forest (5 predictors considered at each leaf split, at least 10 observation per node, bootstrap sample size of 70%), gradient-boosted trees with low regularization (500 trees, maximum depth of 3 and a learning rate of 0.01), gradient-boosted trees with high regularization: 250 trees, maximum depth of 3 and a learning rate of 0.01, feed-forward neural nets with three hidden layers of size five. For reference, we report two estimators using the full sample: OLS and PDS lasso. We report results for four meta learners: Stacking with CLS, short-stacking with CLS, single best overall and single best by fold. Results are based on 1 000 replications.

Table B.4: Mean bias under non-linear DGP in small samples based on the calibrated Monte Carlo in Section 4.1

	Bootstrap sample size n_b									
	200		400		800		1 600		9 915	
	Bias	s.e.	Bias	s.e.	Bias	s.e.	Bias	s.e.	Bias	s.e.
Full sample estimators:										
OLS	-2339.8	331.6	-2208.9	226.7	-2332.7	165.1	-2440.2	112.5	-2586.2	47.6
PDS-Lasso	-2281.9	329.6	-2740.9	224.9	-2789.1	168.4	-2716.2	114.1	-2597.3	47.7
DDML methods:										
Candidate learners ($K = 10$)										
OLS	-2864.1	370.5	-2607.8	247.4	-2479.6	172.1	-2526.7	115.0	-2606.5	47.8
Lasso with CV (2nd order poly)	-1048.1	357.5	-730.3	232.5	243.2	166.1	467.6	111.8	783.4	43.0
Ridge with CV (2nd order poly)	-1730.2	348.3	-1127.8	232.0	481.5	170.6	833.2	113.3	806.7	43.0
Lasso with CV (10th order poly)	22248.8	25988.5	-2907.1	2774.5	1598.1	3131.5	4485.9	3456.2	185.6	457.2
Ridge with CV (10th order poly)	24105.0	16649.8	2351.6	2060.6	-5341.5	5329.5	-602.5	2313.4	2873.4	552.4
Random forest (low regularization)	-84.4	343.4	18.2	231.6	-5.4	158.3	-89.8	113.2	-63.6	47.7
Random forest (high regularization)	-159.0	333.3	-37.4	223.1	99.8	154.6	-77.1	107.6	-78.1	42.7
Gradient boosting (low regularization)	-281.5	351.7	-32.4	238.7	-11.1	157.8	34.1	108.8	64.6	41.3
Gradient boosting (high regularization)	-89.7	336.4	208.6	229.0	179.2	152.6	182.9	105.2	218.2	41.2
Neural net	1839.2	299.8	2045.7	225.2	1735.2	159.0	209.6	107.6	-472.8	43.6
Meta learners ($K = 10$)										
Stacking: CLS	11936.9	14230.0	-1198.8	965.2	803.6	583.6	915.4	1141.7	437.8	46.8
Stacking: Average	7161.2	5979.0	498.4	566.1	-368.8	751.2	765.1	482.6	451.8	50.7
Stacking: OLS	-2.7e7	2.4e7	-3391.9	2466.9	3033.6	3386.2	6271.2	4906.4	608.4	190.2
Stacking: Single-best	3472.5	4282.0	-385.1	306.5	687.0	541.4	542.5	255.2	154.1	46.1
Short-stacking: CLS	-278.6	311.4	-117.9	208.7	88.3	146.0	188.6	101.9	157.7	41.0
Short-stacking: Average	7161.2	5979.0	498.4	566.1	-368.8	751.2	765.1	482.6	451.8	50.7
Short-stacking: OLS	-524.0	298.3	-209.4	199.9	16.5	142.2	88.7	100.2	126.1	40.8
Short-stacking: Single-best	-602.8	310.3	-363.9	215.6	51.5	150.3	196.2	104.5	66.9	41.3
Pooled stacking: CLS	-382.2	315.4	-319.7	218.0	-73.3	150.7	143.9	103.7	179.9	41.0
Pooled stacking: Average	7161.2	5979.0	498.4	566.1	-368.8	751.2	765.1	482.6	451.8	50.7
Pooled stacking: OLS	-217.6	359.4	-231.1	223.5	-223.3	158.8	-23.4	120.8	152.0	40.9
Pooled stacking: Single-best	-679.0	319.8	-316.5	228.0	-52.0	154.5	159.4	105.9	68.2	41.2
Meta learners ($K = 2$)										
Stacking: CLS	656.2	1786.0	583.5	718.1	-393.0	500.7	-710.3	585.3	133.9	195.6
Stacking: Average	2921.9	1220.8	2639.9	1126.1	318.6	1095.6	-656.9	526.5	94.0	110.3
Stacking: OLS	2.2e7	1.9e7	-8671.2	9731.4	-5593.2	7804.4	971.5	3368.2	-204.8	405.2
Stacking: Single-best	2892.1	1501.3	-442.5	347.7	-387.1	295.4	-311.5	506.8	-121.9	272.9
Short-stacking: CLS	410.7	316.1	-76.2	215.5	-320.8	163.2	-219.0	107.8	162.7	42.0
Short-stacking: Average	2921.9	1220.8	2639.9	1126.1	318.6	1095.6	-656.9	526.5	94.0	110.3
Short-stacking: OLS	-130.7	287.7	-201.9	202.5	-322.2	155.3	-226.9	106.1	123.6	41.7
Short-stacking: Single-best	-28.4	323.0	-363.6	221.3	-546.6	164.0	-320.3	110.4	71.7	42.6
Pooled stacking: CLS	767.3	1131.7	942.5	577.0	-673.3	318.3	-734.7	567.8	209.8	63.7
Pooled stacking: Average	2921.9	1220.8	2639.9	1126.1	318.6	1095.6	-656.9	526.5	94.0	110.3
Pooled stacking: OLS	-1784.7	5455.8	-6585.3	4193.1	-897.4	1388.9	903.3	2480.1	234.5	124.9
Pooled stacking: Single-best	1594.5	1414.5	-26.1	449.7	-557.8	366.5	-176.4	379.9	103.3	45.8

Notes: See Table B.3 notes.

Table B.5: Coverage in small samples based on the calibrated Monte Carlo in Section 4.1

	Bootstrap sample size n_b									
	Panel A. Linear DGP					*Panel B. Non-linear DGP*				
	200	400	800	1600	9915	200	400	800	1 600	9 915
Full sample estimators:										
OLS	0.95	0.95	0.95	0.94	0.95	0.92	0.95	0.94	0.92	0.59
PDS-Lasso	0.96	0.95	0.94	0.94	0.95	0.92	0.95	0.93	0.91	0.59
DDML methods:										
Candidate learners ($K = 10$)										
OLS	0.95	0.95	0.94	0.94	0.94	0.93	0.95	0.94	0.91	0.59
Lasso with CV (2nd order poly)	0.95	0.95	0.94	0.93	0.94	0.94	0.95	0.96	0.94	0.90
Ridge with CV (2nd order poly)	0.96	0.95	0.94	0.94	0.94	0.94	0.96	0.96	0.94	0.90
Lasso with CV (10th order poly)	0.90	0.92	0.91	0.93	0.94	0.87	0.85	0.86	0.89	0.95
Ridge with CV (10th order poly)	0.82	0.87	0.88	0.88	0.95	0.78	0.83	0.84	0.87	0.95
Random forest (low regularization)	0.92	0.93	0.92	0.91	0.91	0.92	0.93	0.94	0.93	0.90
Random forest (high regularization)	0.95	0.94	0.95	0.94	0.95	0.94	0.95	0.96	0.95	0.94
Gradient boosting (low regularization)	0.92	0.93	0.93	0.94	0.95	0.92	0.93	0.95	0.95	0.95
Gradient boosting (high regularization)	0.94	0.95	0.94	0.94	0.95	0.93	0.94	0.97	0.95	0.94
Neural net	0.94	0.92	0.90	0.90	0.95	0.94	0.95	0.95	0.97	0.94
Meta learners ($K = 10$)										
Stacking: CLS	0.94	0.95	0.94	0.93	0.95	0.92	0.94	0.94	0.94	0.95
Stacking: Average	0.94	0.95	0.94	0.94	0.95	0.90	0.93	0.95	0.95	0.95
Stacking: OLS	0.85	0.89	0.89	0.90	0.93	0.82	0.86	0.89	0.92	0.93
Stacking: Single-best	0.94	0.94	0.95	0.94	0.95	0.94	0.94	0.95	0.95	0.94
Short-stacking: CLS	0.95	0.95	0.94	0.94	0.94	0.94	0.96	0.96	0.95	0.94
Short-stacking: Average	0.94	0.95	0.94	0.94	0.95	0.90	0.93	0.95	0.95	0.95
Short-stacking: OLS	0.95	0.94	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.94
Short-stacking: Single-best	0.95	0.95	0.94	0.94	0.94	0.93	0.96	0.96	0.96	0.95
Pooled stacking: CLS	0.95	0.95	0.94	0.94	0.94	0.94	0.95	0.96	0.95	0.94
Pooled stacking: Average	0.94	0.95	0.94	0.94	0.95	0.90	0.93	0.95	0.95	0.95
Pooled stacking: OLS	0.95	0.95	0.94	0.94	0.94	0.94	0.96	0.96	0.95	0.94
Pooled stacking: Single-best	0.95	0.95	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95
Meta learners ($K = 2$)										
Stacking: CLS	0.90	0.90	0.93	0.94	0.94	0.90	0.92	0.93	0.93	0.94
Stacking: Single-best	0.91	0.91	0.93	0.94	0.94	0.92	0.93	0.93	0.93	0.94
Short-stacking: CLS	0.94	0.93	0.94	0.95	0.94	0.93	0.94	0.93	0.94	0.94
Short-stacking: Average	0.90	0.91	0.93	0.93	0.94	0.89	0.90	0.91	0.92	0.94
Short-stacking: OLS	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.95	0.94
Short-stacking: Single-best	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94
Pooled stacking: CLS	0.93	0.90	0.93	0.94	0.95	0.92	0.94	0.93	0.95	0.94
Pooled stacking: Average	0.90	0.91	0.93	0.93	0.94	0.89	0.90	0.91	0.92	0.94
Pooled stacking: OLS	0.88	0.89	0.92	0.93	0.94	0.85	0.90	0.90	0.93	0.94
Pooled stacking: Single-best	0.94	0.92	0.93	0.95	0.94	0.94	0.95	0.94	0.95	0.94

Notes: This table reports coverage of 95% interval estimates in the small sample simulation. See Table B.3 notes for more detail.

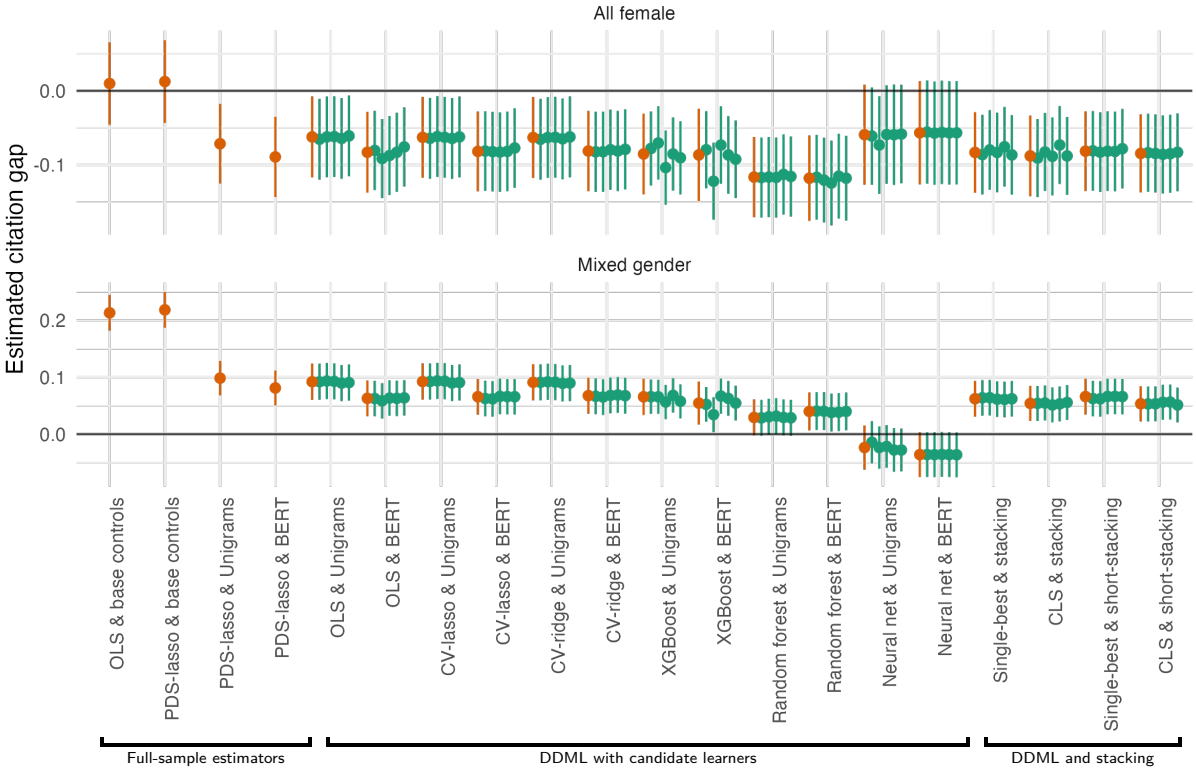
C Gender citation gap

Table C.1: Estimates for the citation penalty of all-female and mixed-gender authored articles

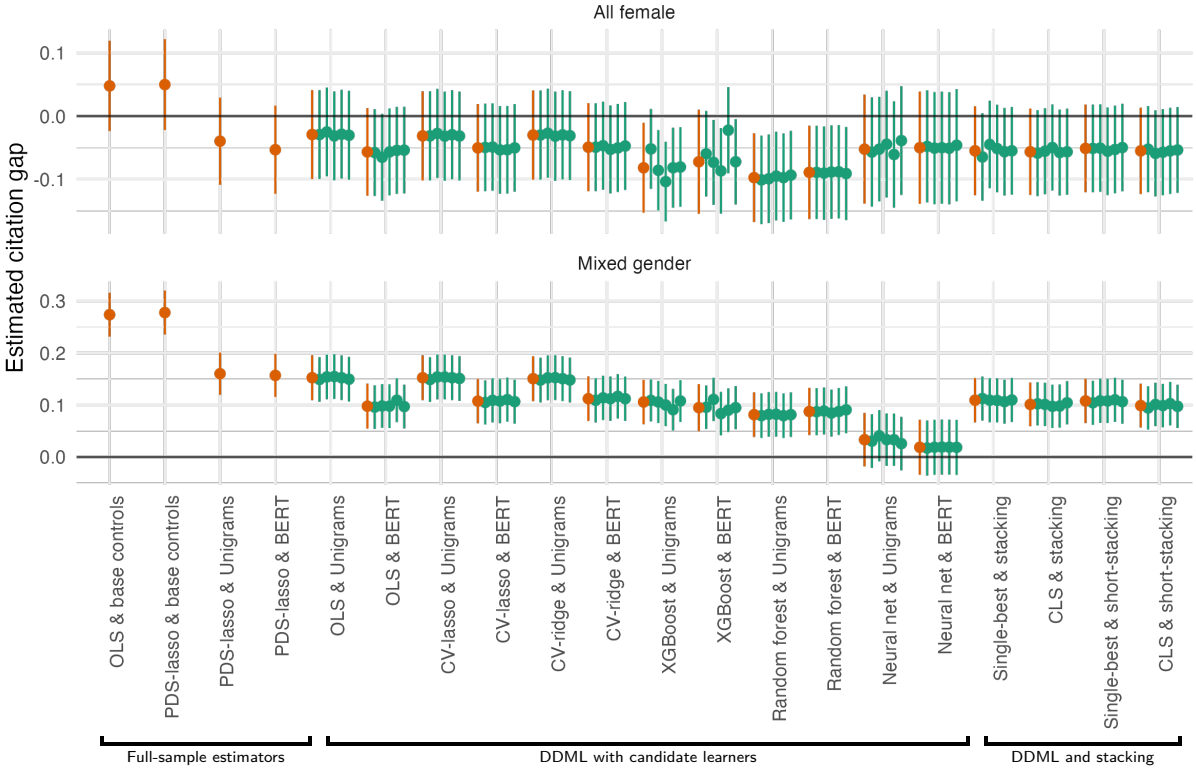
	Log citations		Citation counts	
	All female	*Mixed gender*	*All female*	*Mixed gender*
Panel A. Full-sample estimators				
OLS & base controls	0.022	0.226***	-9.129**	12.168***
	(0.03)	(0.017)	(4.438)	(3.045)
PDS-lasso & base controls	0.025	0.231***	-8.901**	12.317***
	(0.03)	(0.017)	(4.431)	(3.045)
PDS-lasso & Unigrams	-0.056*	0.106***	-15.77***	2.844
	(0.029)	(0.017)	(4.283)	(3.053)
Panel B. DDML with candidate learners				
PDS-lasso & BERT	-0.074**	0.092***	-15.051***	2.125
	(0.029)	(0.017)	(4.428)	(3.053)
OLS & Unigrams	-0.052*	0.103***	-12.293**	2.647
	(0.03)	(0.018)	(6.104)	(3.635)
OLS & BERT	-0.072**	0.066***	-8.686	1.454
	(0.029)	(0.017)	(6.242)	(3.655)
CV-lasso & Unigrams	-0.05*	0.104***	-12.388**	2.648
	(0.03)	(0.018)	(6.116)	(3.621)
CV-lasso & BERT	-0.069**	0.069***	-9.928	2.209
	(0.029)	(0.017)	(6.142)	(3.648)
CV-ridge & Unigrams	-0.051*	0.102***	-12.383**	2.077
	(0.03)	(0.018)	(6.115)	(3.627)
CV-ridge & BERT	-0.067**	0.075***	-9.86	2.056
	(0.029)	(0.017)	(6.155)	(3.636)
XGBoost & Unigrams	-0.065**	0.073***	37.984***	23.481***
	(0.03)	(0.02)	(6.655)	(3.629)
XGBoost & BERT	-0.07**	0.073***	9.473	8.922**
	(0.031)	(0.021)	(7.306)	(3.796)
Random forest & Unigrams	-0.103***	0.044**	0.444	5.255
	(0.029)	(0.017)	(6.098)	(3.623)
Random forest & BERT	-0.106***	0.055***	8.919	10.869***
	(0.031)	(0.018)	(6.148)	(3.667)
Neural net & Unigrams	-0.05	0.005	-16.867***	-9.791***
	(0.036)	(0.021)	(6.317)	(3.647)
Neural net & BERT	-0.047	-0.014	-16.979***	-10.31***
	(0.038)	(0.022)	(6.316)	(3.643)
Panel C. DDML with stacking approaches				
Single-best & stacking	-0.07**	0.072***	-10.204*	2.274
	(0.029)	(0.017)	(6.123)	(3.65)
CLS & stacking	-0.071**	0.063***	-9.485	1.971
	(0.029)	(0.017)	(6.103)	(3.635)
Single-best & short-stacking	-0.069**	0.07***	-9.916	2.254
	(0.029)	(0.017)	(6.143)	(3.643)
CLS & short-stacking	-0.07**	0.062***	-9.529	1.615
	(0.029)	(0.017)	(6.099)	(3.628)

Notes: The table shows median-aggregated estimates of the gender citation gap for all–female and mixed-gender authored articles. We show results using both log citations and citation counts as the outcome variable. Standard errors are robust to heteroskedasticity. See Figure 6 for information on the candidate learners and stacking approaches.

Figure C.1: The citation gap by authors' gender composition



(a) Threshold 60%



(b) Threshold 90%

Notes: The figure shows estimates of θ_0 summarizing average relative difference in total citations between all-male and all-female authorship, and all-male and mixed-gender authorship, respectively using different thresholds for successful classification of an author's sex. See Figure 6 notes for more information.

D Gender wage gap

Table D.1: Stacking weights in the gender wage gap application.

	Conventional stacking			*Short-stacking*			*Mean-squared error*		
	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$
OLS/logit	0.023	0.012	0.242	0.027	0.013	0.211	0.369	0.347	0.161
OLS/logit (simple)	0.004	0.	0.	0.	0.	0.	0.267	0.204	0.223
CV-lasso	0.103	0.136	0.109	0.03	0.076	0.047	0.236	0.178	0.16
CV-ridge	0.189	0.04	0.064	0.225	0.024	0.108	0.237	0.18	0.161
CV-lasso (extended)	0.041	0.157	0.016	0.035	0.266	0.002	0.238	0.18	0.161
CV-ridge (extended)	0.011	0.04	0.011	0.003	0.024	0.022	0.336	0.194	0.161
Random forest 1	0.435	0.506	0.275	0.483	0.507	0.28	0.23	0.176	0.161
Random forest 2	0.	0.	0.	0.	0.	0.	0.258	0.19	0.171
Random forest 3	0.	0.	0.	0.	0.	0.	0.274	0.199	0.179
Gradient boosting 1	0.025	0.008	0.039	0.011	0.003	0.022	0.239	0.183	0.16
Gradient boosting 2	0.15	0.059	0.216	0.175	0.063	0.285	0.254	0.196	0.161
Neural net 1	0.013	0.022	0.	0.	0.	0.	0.349	0.263	0.241
Neural net 2	0.008	0.02	0.027	0.01	0.023	0.023	0.643	0.357	0.176

Notes: The table shows weights of conventional and short-stacking along with the mean-squared prediction error by candidate learners and by variable. The final learner is constrained least squares. The stacking weights are averaged over cross-fitting repetitions. Pooled stacking weights are shown in Appendix Table D.2.

Table D.2: Stacking weights of pooled stacking using constrained least squares.

	Pooled stacking		
	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$
OLS/logit	0.014	0.001	0.257
OLS/logit (simple)	0.	0.	0.
CV-lasso	0.15	0.23	0.136
CV-ridge	0.205	0.064	0.063
CV-lasso (extended)	0.	0.078	0.
CV-ridge (extended)	0.	0.019	0.
Random forest 1	0.462	0.521	0.288
Random forest 2	0.	0.	0.
Random forest 3	0.	0.	0.
Gradient boosting 1	0.	0.	0.008
Gradient boosting 2	0.165	0.071	0.23
Neural net 1	0.	0.	0.
Neural net 2	0.004	0.016	0.018

Notes: The table shows pooled stacking weights for each of the considered candidate learners. The final learner is constrained least squares. The stacking weights are averaged over cross-fitting repetitions.

Table D.3: Stacking weights using single-best final learner.

	Conventional stacking			*Short-stacking*			*Pooled stacking*		
	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$	$g_0(0,X)$	$g_0(1,X)$	$m_0(X)$
OLS/logit	0.	0.	0.	0.	0.	0.	0.	0.	0.
OLS/logit (simple)	0.	0.	0.	0.	0.	0.	0.	0.	0.
CV-lasso	0.06	0.03	0.79	0.	0.	1.	0.	0.	0.4
CV-ridge	0.07	0.	0.04	0.	0.	0.	0.	0.	0.1
CV-lasso (extended)	0.02	0.06	0.03	0.	0.	0.	0.	0.	0.1
CV-ridge (extended)	0.	0.	0.01	0.	0.	0.	0.	0.	0.
Random forest 1	0.85	0.91	0.02	1.	1.	0.	1.	1.	0.
Random forest 2	0.	0.	0.	0.	0.	0.	0.	0.	0.
Random forest 3	0.	0.	0.	0.	0.	0.	0.	0.	0.
Gradient boosting 1	0.	0.	0.11	0.	0.	0.	0.	0.	0.4
Gradient boosting 2	0.	0.	0.	0.	0.	0.	0.	0.	0.
Neural net 1	0.	0.	0.	0.	0.	0.	0.	0.	0.
Neural net 2	0.	0.	0.	0.	0.	0.	0.	0.	0.

Notes: The table shows weights of conventional stacking, short-stacking and pooled stacking by candidate learners and by conditional expectation function. The stacking weights are averaged over cross-fitting repetitions.

Table D.4: Median aggregate estimates by stacking approach and by final learner

	Final learner			
	Unweighted average	*CLS*	*OLS*	*Single-best*
Regular stacking	-0.101	-0.075	-0.197	-0.061
	(0.017)*	(0.028)*	(11.894)	(0.069)
Short- stacking	-0.101	-0.076	-0.001	-0.085
	(0.017)*	(0.028)*	(0.184)	(0.065)

Notes: The table reports median aggregate estimates by stacking type and final learner. See Figure 7 for more information.

Table D.5: Median aggregate estimates for each candidate learner

	Gender wage gap
OLS/logit	-0.12
	(0.094)
CV-lasso	-0.067
	(0.063)
CV-ridge	-0.064
	(0.09)
OLS/logit (simple)	-0.12
	(0.016)*
CV-lasso (extended)	-0.055
	(0.076)
CV-ridge (extended)	-0.173
	(0.19)
Random forest 1	-0.079
	(0.023)*
Random forest 2	-0.105
	(0.016)*
Random forest 3	-0.11
	(0.015)*
Gradient boosting 1	-0.075
Observations	4836

Notes: The table reports median aggregate estimates by candidate learner. See Figure 7 for more information.