nature portfolio

https://doi.org/10.1038/s42256-022-00470-y

Supplementary information

Automated causal inference in application to randomized controlled clinical trials

In the format provided by the authors and unedited

Supplementary material

Ji Q. Wu[1,*], Nanda Horeweg[2], Marco de Bruyn[3], Remi A. Nout[2,], Ina M. Jürgenliemk-Schulz[4], Ludy C.H.W. Lutgens[5], Jan J. Jobsen[6,***], Elzbieta M. van der Steen-Banasik[7], Hans W. Nijman[3], Vincent T.H.B.M. Smit[8], Tjalling Bosse[8], Carien L. Creutzberg[2], and Viktor H. Koelzer[1,*]**

[1] Department of Pathology and Molecular Pathology, University Hospital, University of Zurich, Zurich, Switzerland.
[2] Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands.
[3] Department of Obstetrics and Gynecology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
[4] Department of Radiation Oncology, University Medical Center Utrecht, Utrecht, The Netherlands.
[5] Maastricht Radiation Oncology Clinic, Maastricht, The Netherlands.
[6] Department of Radiotherapy, Medisch Spectrum Twente, Enschede, The Netherlands.
[7] Radiotherapiegroep, Arnhem, The Netherlands.
[8] Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.
[*] Corresponding authors, Jiqing.Wu@usz.ch, Viktor.Koelzer@usz.ch
[**] Currently employed at department of Radiotherapy, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands.
[***] Currently employed at department of Clinical Epidemiology, Medisch Spectrum Twente, Enschede, The Netherlands.

ABSTRACT

Randomized controlled trials (RCTs) are considered as the gold standard for testing causal hypotheses in the clinical domain. However, the investigation of prognostic variables of patient outcome in a hypothesized cause-effect route is not feasible using standard statistical methods. Here, we propose a new automated causal inference method (AutoCI) built upon the invariant causal prediction (ICP) framework for the causal re-interpretation of clinical trial data. Compared to existing methods, we show that the proposed AutoCI allows to efficiently determine the causal variables with a clear differentiation on two real-world RCTs of endometrial cancer patients with mature outcome and extensive clinicopathological and molecular data. This is achieved via suppressing the causal probability of non-causal variables by a wide margin. In ablation studies, we further demonstrate that the assignment of causal probabilities by AutoCI remain consistent in the presence of confounders. In conclusion, these results confirm the robustness and feasibility of AutoCI for future applications in real-world clinical analysis.

Abbreviation	Definition
RCT	Randomised controlled trials
PORTEC	Post operative radiation therapy in endometrial carcinoma
EC	Endometrial carcinoma (cancer)
ESGO	European Society of Gynaecological Oncology
ESTRO	European Society for Radiotherapy and Oncology
ESMO	European Society of Medical Oncology
Grade	Tumor grading
LVSI	Lymphovascular space invasion
POLEmut	Polymerase epsilon mutant EC
MMRd	Mismatch repair deficient EC
p53abn	p53 abnormal EC
NSMP	EC with no specific molecular profile
L1CAM	L1 cell adhesion molecule
P	Pathological variables
PM	Pathological and molecular variables
PMI	Pathological, molecular and immune variables
HR	Hazard ratio
CI	Confidence interval
i.d.	identically distributed
NP	Non-deterministic polynomial-time
SCM	Structural causal model
ICP	Invariant causal prediction
PRED	Predicate module
FID	Fréchet inception distance
JS	Jaccard similarity
FWER	Family-wise error rate
ABCD	Active budgeted causal design strategy

Table 1. The abbreviation table of clinical, statistical and causal definitions.

References

1. Gaunt, A. L., Brockschmidt, M., Kushman, N. & Tarlow, D. Differentiable programs with neural libraries. *Int. Conf. on Mach. Learn.* 1213–1222 (2017).

2. Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B. & Wu, J. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *Int. Conf. on Learn. Represent.* (2018).

3. Vedantam, R. *et al.* Probabilistic neural symbolic models for interpretable visual question answering. *Int. Conf. on Mach. Learn.* 6428–6437 (2019).

4. Ellis, K. *et al.* Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381* (2020).

5. Valkov, L., Chaudhari, D., Srivastava, A., Sutton, C. & Chaudhuri, S. Houdini: Lifelong learning as program synthesis. *Adv. Neural Inf. Process. Syst.* 8687–8698 (2018).

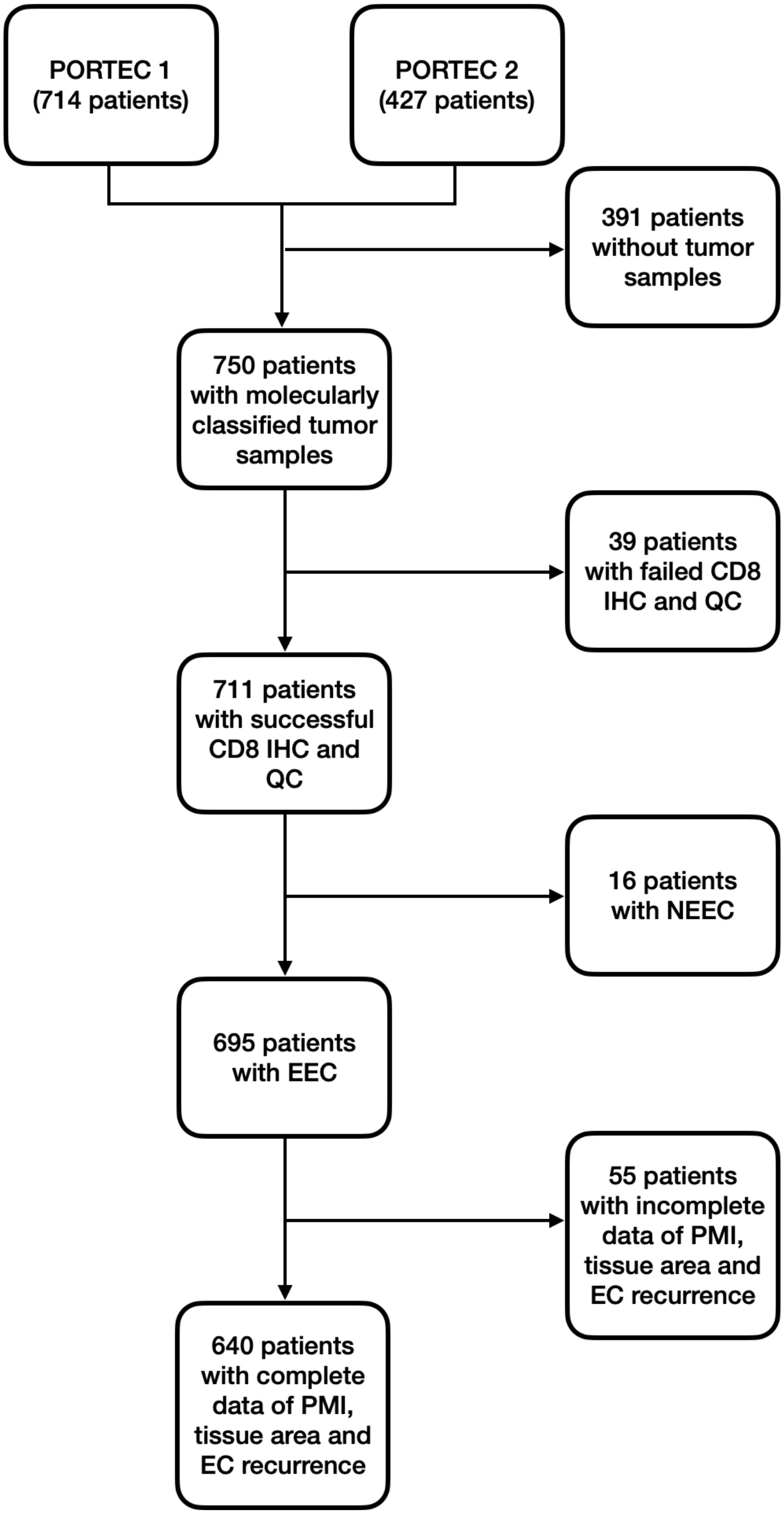
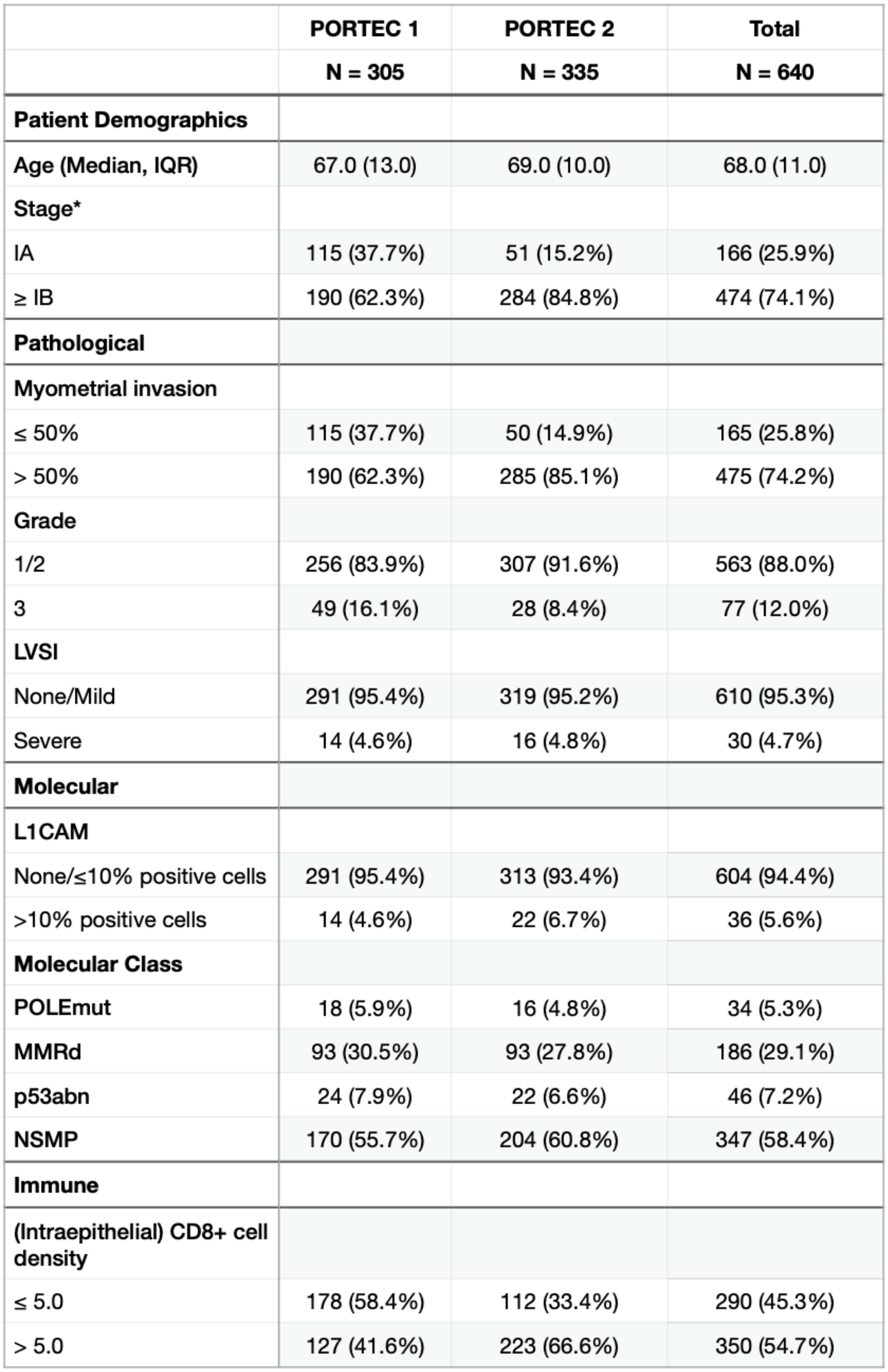

PORTEC 1
(714 patients)
PORTEC 2
(427 patients)
391 patients without tumor samples
750 patients with molecularly classified tumor samples
39 patients with failed CD8 IHC and QC
711 patients with successful CD8 IHC and QC
16 patients with NEEC
695 patients with EEC
55 patients with incomplete data of PMI, tissue area and EC recurrence
640 patients with complete data of PMI, tissue area and EC recurrence

Figure 1. The consort diagram presenting the process of patient selection. Abbreviations: QC - quality control, IHC - immunohistochemistry, EEC- endometrioid endometrial carcinoma, NEEC- non-endometrioid endometrial carcinoma.

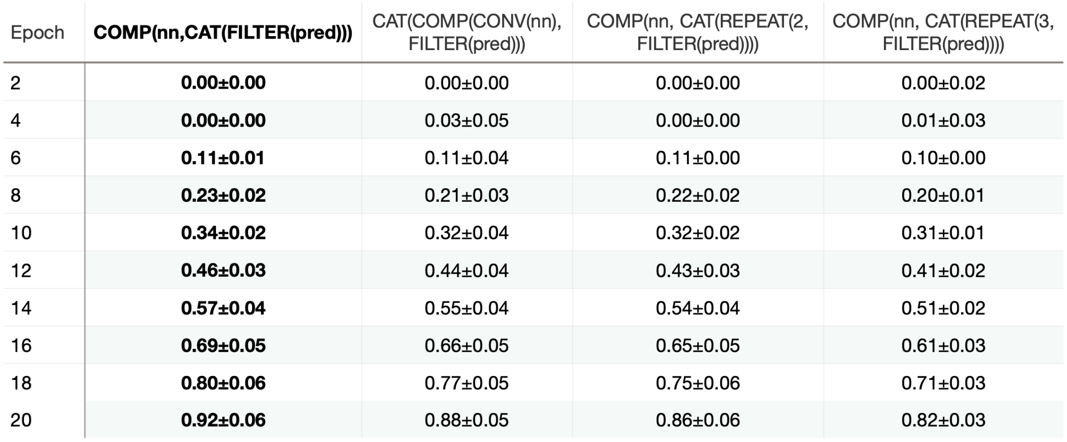
	Excluded	**Included**	
Characteristics	**N = 501**	**N = 640**	**p-value**
Age (Median, IQR)	67.0 (13.0)	68.0 (11.0)	< 0.001
Stage*			
IA	199 (39.7%)	166 (25.9%)	< 0.0001
≥ IB	302 (60.3%)	474 (74.1%)	
Myometrial invasion			
≤ 50%	200 (39.9%)	165 (25.8%)	< 0.0001
> 50%	301 (60.1%)	475 (74.2%)	
Grade			
1/2	426 (85.0%)	563 (88.0%)	0.15
3	75 (15.0%)	77 (12.0%)	
LVSI			
None/Mild	300 (94.9%)	610 (95.3%)	0.80
Severe	16 (5.1%)	30 (4.7%)	
Received adjuvant treatment			
None	212 (42.3%)	160 (25.0%)	<0.0001
Vaginal brachytherapy	240 (47.9%)	314 (49.1%)	
Pelvic EBRT	49 (9.8%)	166 (25.9%)	
Recurrence free survival**			
Mean RFS (years, SE, 95% CI)	14.81 (0.25, 14.32-15.31)	15.24 (0.228, 14.70-15.68)	0.140
5-year RFS (%, SE)	85.7 (0.016)	89.8 (0.012)	
Overall survival**			
Mean OS (years, SE, 95% CI)	12.54 (0.33, 11.89-13.20)	12.60 (0.25, 12.12-13.09)	0.450
5-year OS (%, SE)	81.2 (0.017)	85.7 (0.014)	

Table 2. The characteristics comparison of excluded and included patients. * After a posteriori central review 2 cases were classified as stage II, 2 as stage IIIA and 1 as stage IIIB. ** The p-values of RFS and OS are computed from log-rank test.



	PORTEC 1	PORTEC 2	Total
	N = 305	N = 335	N = 640
Patient Demographics			
Age (Median, IQR)	67.0 (13.0)	69.0 (10.0)	68.0 (11.0)
Stage*			
IA	115 (37.7%)	51 (15.2%)	166 (25.9%)
≥ IB	190 (62.3%)	284 (84.8%)	474 (74.1%)
Pathological			
Myometrial invasion			
≤ 50%	115 (37.7%)	50 (14.9%)	165 (25.8%)
> 50%	190 (62.3%)	285 (85.1%)	475 (74.2%)
Grade			
1/2	256 (83.9%)	307 (91.6%)	563 (88.0%)
3	49 (16.1%)	28 (8.4%)	77 (12.0%)
LVSI			
None/Mild	291 (95.4%)	319 (95.2%)	610 (95.3%)
Severe	14 (4.6%)	16 (4.8%)	30 (4.7%)
Molecular			
L1CAM			
None/≤10% positive cells	291 (95.4%)	313 (93.4%)	604 (94.4%)
>10% positive cells	14 (4.6%)	22 (6.7%)	36 (5.6%)
Molecular Class			
POLEmut	18 (5.9%)	16 (4.8%)	34 (5.3%)
MMRd	93 (30.5%)	93 (27.8%)	186 (29.1%)
p53abn	24 (7.9%)	22 (6.6%)	46 (7.2%)
NSMP	170 (55.7%)	204 (60.8%)	347 (58.4%)
Immune			
(Intraepithelial) CD8+ cell density			
≤ 5.0	178 (58.4%)	112 (33.4%)	290 (45.3%)
> 5.0	127 (41.6%)	223 (66.6%)	350 (54.7%)

Table 3. The characteristics of study participants for PORTEC 1 and 2. * After a posteriori central review 2 cases were classified as stage II, 2 as stage IIIA and 1 as stage IIIB.

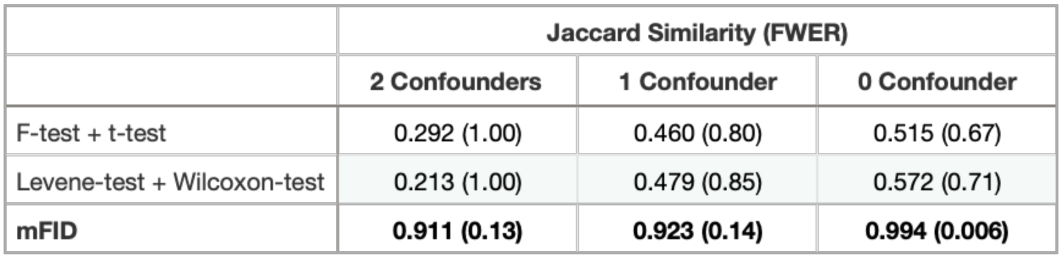


Epoch	**COMP(nn,CAT(FILTER(pred)))**	CAT(COMP(CONV(nn), FILTER(pred)))	COMP(nn, CAT(REPEAT(2, FILTER(pred))))	COMP(nn, CAT(REPEAT(3, FILTER(pred))))
2	**0.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.02
4	**0.00±0.00**	0.03±0.05	0.00±0.00	0.01±0.03
6	**0.11±0.01**	0.11±0.04	0.11±0.00	0.10±0.00
8	**0.23±0.02**	0.21±0.03	0.22±0.02	0.20±0.01
10	**0.34±0.02**	0.32±0.04	0.32±0.02	0.31±0.01
12	**0.46±0.03**	0.44±0.04	0.43±0.03	0.41±0.02
14	**0.57±0.04**	0.55±0.04	0.54±0.04	0.51±0.02
16	**0.69±0.05**	0.66±0.05	0.65±0.05	0.61±0.03
18	**0.80±0.06**	0.77±0.05	0.75±0.06	0.71±0.03
20	**0.92±0.06**	0.88±0.05	0.86±0.06	0.82±0.03

Table 4. The JS score and its standard deviation of compared type-safe candidates for PORTEC study (PMI).

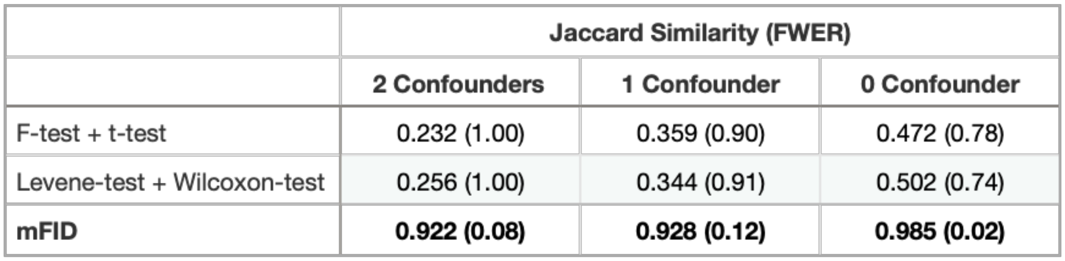
	Task	Typed	Functional	Code availability
NTPT[1]	Misc.	✗	✗	✗
NS-CL[2]	VQA	✗	✓	✓
Prob-NMN[3]	VQA	✗	✓	✓
DreamCoder[4]	Misc.	✓	✓	✗
HOUDINI[5]	**Misc.**	✓	✓	✓

Table 5. The comparison between existing program synthesis languages.



	Jaccard Similarity (FWER)		
	2 Confounders	**1 Confounder**	**0 Confounder**
F-test + t-test	0.292 (1.00)	0.460 (0.80)	0.515 (0.67)
Levene-test + Wilcoxon-test	0.213 (1.00)	0.479 (0.85)	0.572 (0.71)
mFID	**0.911 (0.13)**	**0.923 (0.14)**	**0.994 (0.006)**

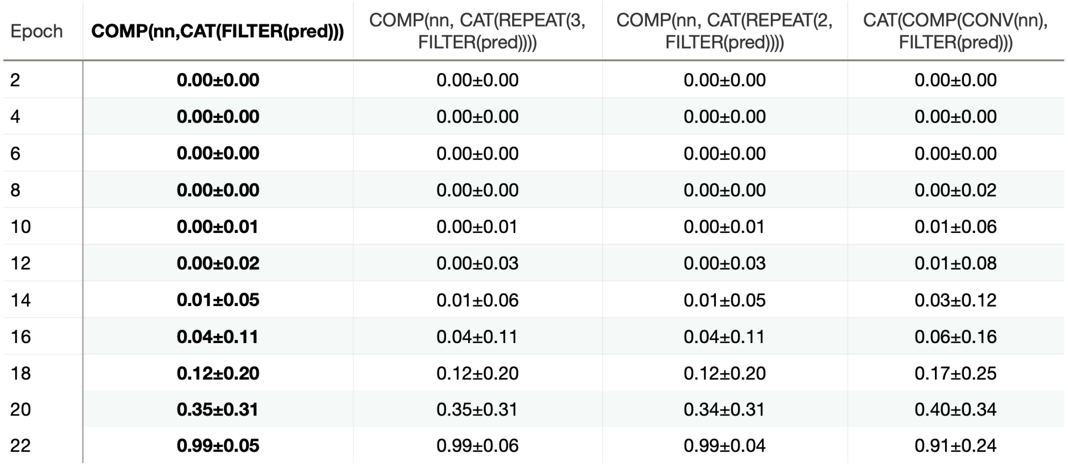
Finite sample setting



	Jaccard Similarity (FWER)		
	2 Confounders	**1 Confounder**	**0 Confounder**
F-test + t-test	0.232 (1.00)	0.359 (0.90)	0.472 (0.78)
Levene-test + Wilcoxon-test	0.256 (1.00)	0.344 (0.91)	0.502 (0.74)
mFID	**0.922 (0.08)**	**0.928 (0.12)**	**0.985 (0.02)**

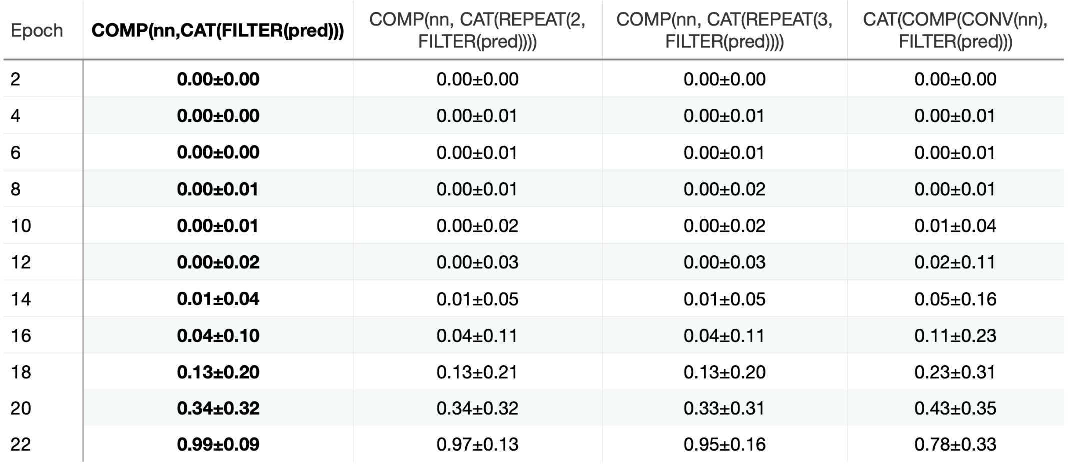
ABCD setting

Table 6. The comparison of statistical measurements for toy experiments. Top: The results of the compared statistical measurements for the Finite sample setting. Bottom: The results of the compared statistical measurements for the ABCD setting. Here, all the measurements are applied for training the same type-safe function COMP(nn, CAT(FILTER(pred))) under the proposed causal differentiable learning scheme. F-test + t-test is used in ICP and AICP. Levene-test + Wilcoxon-test is used in NICP.



Epoch	**COMP(nn,CAT(FILTER(pred)))**	COMP(nn, CAT(REPEAT(3, FILTER(pred))))	COMP(nn, CAT(REPEAT(2, FILTER(pred))))	CAT(COMP(CONV(nn), FILTER(pred)))
2	**0.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
4	**0.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
6	**0.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
8	**0.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.02
10	**0.00±0.01**	0.00±0.01	0.00±0.01	0.01±0.06
12	**0.00±0.02**	0.00±0.03	0.00±0.03	0.01±0.08
14	**0.01±0.05**	0.01±0.06	0.01±0.05	0.03±0.12
16	**0.04±0.11**	0.04±0.11	0.04±0.11	0.06±0.16
18	**0.12±0.20**	0.12±0.20	0.12±0.20	0.17±0.25
20	**0.35±0.31**	0.35±0.31	0.34±0.31	0.40±0.34
22	**0.99±0.05**	0.99±0.06	0.99±0.04	0.91±0.24

Table 7. The JS score and its standard deviation of compared type-safe candidates for toy experiments (Finite sample setting).



Epoch	**COMP(nn,CAT(FILTER(pred)))**	COMP(nn, CAT(REPEAT(2, FILTER(pred))))	COMP(nn, CAT(REPEAT(3, FILTER(pred))))	CAT(COMP(CONV(nn), FILTER(pred)))
2	**0.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
4	**0.00±0.00**	0.00±0.01	0.00±0.01	0.00±0.01
6	**0.00±0.00**	0.00±0.01	0.00±0.01	0.00±0.01
8	**0.00±0.01**	0.00±0.01	0.00±0.02	0.00±0.01
10	**0.00±0.01**	0.00±0.02	0.00±0.02	0.01±0.04
12	**0.00±0.02**	0.00±0.03	0.00±0.03	0.02±0.11
14	**0.01±0.04**	0.01±0.05	0.01±0.05	0.05±0.16
16	**0.04±0.10**	0.04±0.11	0.04±0.11	0.11±0.23
18	**0.13±0.20**	0.13±0.21	0.13±0.20	0.23±0.31
20	**0.34±0.32**	0.34±0.32	0.33±0.31	0.43±0.35
22	**0.99±0.09**	0.97±0.13	0.95±0.16	0.78±0.33

Table 8. The JS score and its standard deviation of compared type-safe candidates for toy experiments (ABCD setting).